Word-Level Contextual Sentiment Analysis with Interpretability

Tomoki Ito,[1] Kota Tsubouchi,[2] Hiroki Sakaji,[1] Tatsuo Yamashita,[2] Kiyoshi Izumi[1]
[1]Graduate School of Engineering, The University of Tokyo, [2]Yahoo Japan Corporation
m2015titoh@socsim.org, {ktsubouc, tayamash}@yahoo-corp.jp, {sakaji, izumi}@sys.t.u-tokyo.ac.jp

Abstract

Word-level contextual sentiment analysis (WCSA) is an important task for mining reviews or opinions. When analyzing this type of sentiment in the industry, both the interpretability and practicality are often required. However, such a WCSA method has not been established. This study aims to develop a WCSA method with interpretability and practicality. To achieve this aim, we propose a novel neural network architecture called Sentiment Interpretable Neural Network (SINN). To realize this SINN practically, we propose a novel learning strategy called Lexical Initialization Learning (LEXIL). SINN is interpretable because it can extract word-level contextual sentiment through extracting word-level original sentiment and its local and global word-level contexts. Moreover, LEXIL can develop the SINN without any specific knowledge for context; therefore, this strategy is practical. Using real textual datasets, we experimentally demonstrate that the proposed LEXIL is effective for improving the interpretability of SINN and that the SINN features both the high WCSA ability and high interpretability.

1 Introduction

1.1 Motivation

Word-level Contextual Sentiment Analysis (WCSA) is the task for assigning word-level sentiment score to each term in a review by considering the contextual influence to it from the other terms. For example, "good" originally has a positive meaning. However, in the phrase "not good", this word is shifted by "not" and its sentiment becomes negative. This WCSA is known to be valuable for mining reviews or opinions (Wilson, Wiebe, and Hoffman 2005) because pinpointing positive or negative expressions as shown in the sentences below is often required in the industry.

(1) In total, we are in a $bull^+$ market.

(2) This room is not $clean^-$.

(3) Products in this shop are too $expensive^-$.

By pinpointing positive or negative expressions, we can identify the detailed positive or negative attitude of consumers. For example, from the third review listed above, we see that the problem for this shop is caused by price, therefore, the price should be improved.

Several studies conducted on this topic; however, most lack practicality or interpretability, which is a crucial problem in industrial usage. Methods using annotated datasets for word-level contextual sentiments (Mohammad, Kiritchenko, and Zhu 2013; Nakov et al. 2013; Rosenthal et al. 2014; Schulder et al. 2017) or specific knowledge (Li et al. 2013; Wilson, Wiebe, and Hoffman 2005; Kiritchenko and Mohammad 2016) have been proposed. However, these approaches are not practical because such annotated datasets or specific knowledge are typically not available for analyzing specialized documents (e.g., legal or financial documents) or minor languages. Other approaches leveraging interpretation through neural networks (Bach et al. 2017; Sundararajan, Taly, and Yan 2017; Karen, Andrea, and Zisserman 2013; Springenberg et al. 2015) were developed, which require only reviews and their document-level sentiment tags. These methods are practical because document-level tags are more available than contextual word-level or phrase tags. However, they lack interpretability because they cannot explain the process of the analysis; therefore, they cannot be applied in cases where explanations are required.

1.2 Purpose

In response, this study aims to develop a WCSA method with both the *interpretability* and *practicality*.

Interpretability To satisfy the interpretability, we aim to develop a WCSA method that can extract word-level contextual sentiment in a review through extracting the following word-level original sentiment, local word-level contexts, and global word-level contexts, as shown in Figure 1.

1) Word-level original sentiment represents the sentiment of each word where it originally has (e.g., scores in a word sentiment dictionary (Hutto and Gilbert 2014)).

2) Local word-level context represents whether each term in a review is shifted or not by the contexts of multiple words or phrases (e.g., "up" in "did not go up." and "bullish" in "manipulate bullish opinion on the stock market").

3) Global word-level context represents the important part of an entire review.

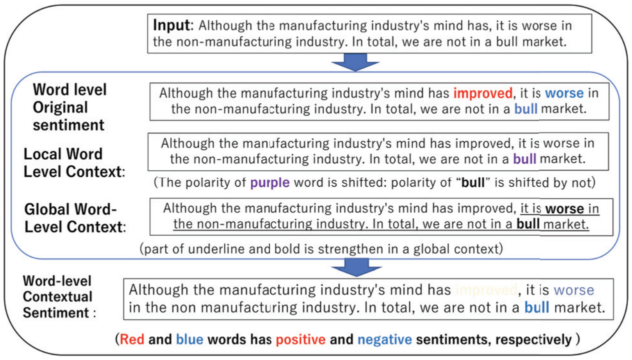
Input: Although the manufacturing industry's mind has, it is worse in the non-manufacturing industry. In total, we are not in a bull market.
Word level Original sentiment
Although the manufacturing industry's mind has improved, it is worse in the non-manufacturing industry. In total, we are not in a bull market.
Local Word Level Context:
Although the manufacturing industry's mind has improved, it is worse in the non-manufacturing industry. In total, we are not in a bull market.
(The polarity of purple word is shifted: polarity of "bull" is shifted by not)
Global Word-Level Context:
Although the manufacturing industry's mind has improved, it is worse in the non-manufacturing industry. In total, we are not in a bull market.
(part of underline and bold is strengthen in a global context)
Word-level Contextual Sentiment :
Although the manufacturing industry's mind has improved, it is worse in the non manufacturing industry. In total, we are not in a bull market.
(Red and blue words has positive and negative sentiments, respectively)

Figure 1: Goal. Word-level contextual sentiment analysis (WCSA) through assigning three types of scores

Humans tend to judge word-level contextual sentiment by these three scores (Li et al. 2013; Schulder et al. 2017). Thus, this type of explanation should be agreeable for humans, and valuable in a situation where explanations are required.

Practicality To satisfy the practicality, we aim to develop a WCSA model using (1) a large number of reviews with their positive or negative sentiment tags, and (2) a small word sentiment dictionary composed of a few hundred word-level original sentiment scores, following the previous work for word-level sentiment analysis (Ito et al. 2018). We do not use any contextual word or phrase-level tags, or specific knowledge for word-level contexts.

1.3 Approach

To achieve this aim, we propose a novel neural network architecture called Sentiment Interpretable NN (*SINN*). Moreover, to realize this SINN practically, we propose a novel learning strategy called Lexical Initialization Learning (*LEXIL*).

SINN The proposed SINN has the following four interpretable layers: word-level original sentiment layer (*WOSL*), local word-level context layer (*LWCL*), global word-level context layer (*GWCL*), and word-level contextual sentiment layer (*WCSL*), which represent the word-level original sentiment, local word-level context, global word-level context, and word-level contextual sentiment, respectively. WOSL is represented in a dictionary-like manner. LWCL and GWCL are represented using long short-term memory (LSTM) cells (Schuster and Paliwal 1997). The values of WCSL are represented by multiplying the values of these three layers. These four layers enable the SINN to explain its analysis result in an agreeable manner by humans (= *interpretable*).

LEXIL In developing the SINN, the realization of the interpretability in the layers is a crucial problem because general back-propagation techniques with reviews and their sentiment tags cannot realize this interpretability. To solve this problem, we propose a novel learning strategy called LEXical Initialization Learning (LEXIL). In LEXIL, the values of the WOSL are initialized using a small word sentiment dictionary, and this initialization leads to improving the interpretability in WOSL, LWCL, and GWCL. LEXIL can develop the SINN with only (1) a large number of reviews with their positive or negative sentiment tags, and (2) a few hundred word sentiment scores initially obtained from a sentiment dictionary (a very small resource), and it does not require any specific knowledge for context. Thus, this LEXIL *should be practical*.

1.4 Contribution

Our contributions are summarized as follows

(1) We propose a novel NN architecture called SINN that can analyze word-level contextual sentiment through explaining its analysis result (Section 2).

(2) To realize this SINN practically, we propose a novel learning strategy called Lexical Initialization learning (LEXIL).

(3) We experimentally demonstrate that (a) LEXIL improves the interpretability of the layers in SINN and that (b) the SINN has both the high WCSA ability and high interpretability. In most cases, the SINN outperformed the other state-of-the-art methods in WCSA task, even though it can explain the analysis result in an interpretable form.

2 SINN

This section describes the proposed SINN. A SINN can be developed through LEXIL (Section 2.3) using a training dataset $\{(\mathbf{Q}_n, d^{\mathbf{Q}_n})\}_{n=1}^{N}$, and a small word sentiment dictionary. Here, N is the training data size, $\mathbf{Q}_n$ is a review, and $d^{\mathbf{Q}_n}$ is its sentiment tag (1 is positive and 0 is negative). Assume that each review $\mathbf{Q}_n$ has L sentences and each sentence contains T words. $w_{it}^{\mathbf{Q}_n}$ represents the tth word in the ith sentence. After the SINN has been developed, it can analyze word-level contextual sentiment with explaining its analysis result, as shown in Figure 2.

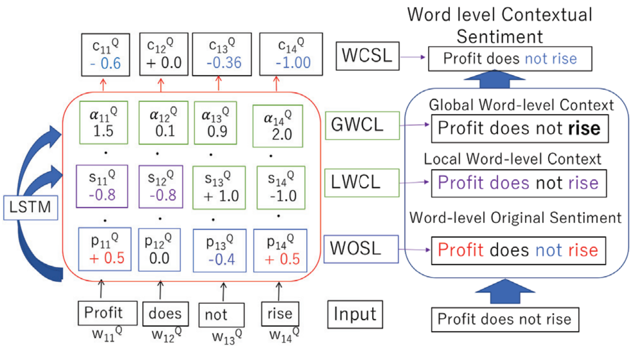

Figure 2: The architecture of the SINN

2.1 Structure of the SINN

This section explains the structure of SINN that includes the WOSL, LWC, GWCL and WCSL, as shown in Figure 2.

Notation We first define several symbols. Let $\{w_i\}_{i=1}^{v}$ be the terms that appear in a text corpus, v be the vocabulary size, and $I(w_i)$ be the vocabulary index of word w_i where $I(w_i) = i$. Let $\boldsymbol{w}_i^{em} \in \mathbb{R}^e$ be an embedding representation of word w_i where $\|\boldsymbol{w}_i^{em}\|_2 = 1$, and the embedding matrix $\boldsymbol{W}^{em} \in \mathbb{R}^{v \times e}$ be $[\boldsymbol{w}_1^{em^T}, \cdots, \boldsymbol{w}_v^{em^T}]^T$ where e is the dimension size of the word embeddings. $\boldsymbol{W}^{em}$ is constant and obtained using the skip-gram method (Mikolov et al. 2013) and the text corpus in a training dataset.

WOSL Given a review $\mathbf{Q} = \{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$, this layer converts the words $\{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ to word-level original sentiment representations $\{\{p_{it}^{\mathbf{Q}}\}_{t=1}^{n}\}_{i=1}^{L}$ as

$$p_{it}^{\mathbf{Q}} = w^{p}_{I(w_{it}^{\mathbf{Q}})} \tag{1}$$

where $\boldsymbol{W}^p \in \mathbb{R}^v$ represents the original sentiment scores of words, and w_i^p is the ith element of $\boldsymbol{W}^p$. The w_i^p value corresponds to the original sentiment score of the word w_i.

LWCL converts words $\{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ to their embeddings $\{\{\boldsymbol{e}_{it}^{\mathbf{Q}}\}_{t=1}^{L}\}_{i=1}^{L}$ using $\boldsymbol{W}^{em}$, and converts them to context representations $\{\{\overrightarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ and $\{\{\overleftarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ using forward and backward LSTMs, $\overrightarrow{\mathrm{LSTM}}$ and $\overleftarrow{\mathrm{LSTM}}$: $\overrightarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}} = \overrightarrow{\mathrm{LSTM}}(\boldsymbol{e}_{it}^{\mathbf{Q}}), \overleftarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}} = \overleftarrow{\mathrm{LSTM}}(\boldsymbol{e}_{it}^{\mathbf{Q}})$.

Then, it converts them to right and left oriented sentiment shift representations, $\overrightarrow{s}_{it}^{\mathbf{Q}}$ and $\overleftarrow{s}_{it}^{\mathbf{Q}}$:

$\overleftarrow{s}_{it}^{\mathbf{Q}} := \tanh(\boldsymbol{v}^{left^T} \overleftarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}}), \overrightarrow{s}_{it}^{\mathbf{Q}} := \tanh(\boldsymbol{v}^{right^T} \overrightarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}})$.

Here, $\boldsymbol{v}^{right}$ and $\boldsymbol{v}^{left} \in \mathbb{R}^e$ are parameter values. $\overrightarrow{s}_{it}^{\mathbf{Q}}$ and $\overleftarrow{s}_{it}^{\mathbf{Q}}$ denote whether the sentiment of $w_{it}^{\mathbf{Q}}$ is shifted or not by the left-side and right-side terms of $w_{it}^{\mathbf{Q}}$: $\{w_{it'}^{\mathbf{Q}}\}_{t'=1}^{t-1}$ and $\{w_{it'}^{\mathbf{Q}}\}_{t'=t+1}^{T}$, respectively. Finally, $\overrightarrow{s}_{it}^{\mathbf{Q}}$ and $\overleftarrow{s}_{it}^{\mathbf{Q}}$ are converted into word-level sentiment shift scores $s_{it}^{\mathbf{Q}}$:

$$s_{it}^{\mathbf{Q}} := \overrightarrow{s}_{it}^{\mathbf{Q}} \cdot \overleftarrow{s}_{it}^{\mathbf{Q}}. \tag{2}$$

where $s_{it}^{\mathbf{Q}}$ denotes whether the sentiment of $w_{it}^{\mathbf{Q}}$ is shifted ($s_{it}^{\mathbf{Q}} < 0$) or not ($s_{it}^{\mathbf{Q}} \geq 0$).

GWCL This layer converts terms $\{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ into the global word-level context scores $\{\{\alpha_{it}\}_{t=1}^{T}\}_{i=1}^{L}$. First, using a revised self-attention mechanism (Vaswani et al. 2016), the word-level attention scores are represented as

$$\beta_{it}^{\mathbf{Q}} := \sum_{t'=1}^{T} \frac{e^{\tanh(\overrightarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}^T} \overrightarrow{\boldsymbol{h}}_{it'}^{\mathbf{Q}} + \overleftarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}^T} \overleftarrow{\boldsymbol{h}}_{it'}^{\mathbf{Q}})}}{\sum_{t=1}^{T} e^{\tanh(\overrightarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}^T} \overrightarrow{\boldsymbol{h}}_{it'}^{\mathbf{Q}} + \overleftarrow{\boldsymbol{h}}_{it}^{\mathbf{Q}^T} \overleftarrow{\boldsymbol{h}}_{it'}^{\mathbf{Q}})}}. \tag{3}$$

Using the sentence-level attention mechanism (Yang et al. 2016), the sentence-level attention scores are represented as

$$\beta_{i}^{\mathbf{Q}} = \frac{e^{AttRNN(\{\boldsymbol{e}_{it}^{\mathbf{Q}}\}_{t=1}^{T})^T \boldsymbol{v}^s}}{\sum_{i=1}^{L} e^{AttRNN(\{\boldsymbol{e}_{it}^{\mathbf{Q}}\}_{t=1}^{T})^T \boldsymbol{v}^s}} \tag{4}$$

where $AttRNN(\cdot)$ is a sentence level context vector produced by the word-level Attention RNN (Yang et al. 2016).

Using these two attention scores, the global word-level context scores are represented by following

$$\alpha_{it}^{\mathbf{Q}} := \beta_{it}^{\mathbf{Q}} \cdot \beta_{t}^{\mathbf{Q}} \tag{5}$$

WCSL represents the word-level contextual sentiment scores $\{\{c_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ using the WOSL, LWCL and GWCL:

$$c_{it}^{\mathbf{Q}} := p_{it}^{\mathbf{Q}} \cdot s_{it}^{\mathbf{Q}} \cdot \alpha_{it}^{\mathbf{Q}}. \tag{6}$$

2.2 Key Idea of LEXIL

In realizing the interpretability of SINN, the realization of the interpretability in WOSL and LWCL is especially difficult. Through the learning with Lo (will be defined in Section 2.3), the WCSL learns to represent word-level contextual sentiment. However, this learning strategy alone cannot realize the interpretability in WOSL and LWCL.

For example, if the word-level contextual sentiment of term $w_{it}^{\mathbf{Q}}$ $(= c_{it}^{\mathbf{Q}})$ is accurately negative, the following two cases are possible: (1) the word-level original sentiment of $w_{it}^{\mathbf{Q}}$ is positive ($p_{it}^{\mathbf{Q}} > 0$) and $w_{it}^{\mathbf{Q}}$ is shifted ($s_{it}^{\mathbf{Q}} < 0$), or (2) the word-level original sentiment of $w_{it}^{\mathbf{Q}}$ is negative ($p_{it}^{\mathbf{Q}} < 0$) and $w_{it}^{\mathbf{Q}}$ is not shifted ($s_{it}^{\mathbf{Q}} > 0$). In general learning, WOSL and LWCL cannot choose the accurate case automatically.

We assume that this problem can be solved by initially limiting the polarity of $p_{it}^{\mathbf{Q}}$ to the accurate case for a few words because this limitation leads to the accurate choice from the above two cases. Therefore, this type of limitation can lead to the learning of $s_{it}^{\mathbf{Q}}$ within the appropriate case. It is assumed that the effect of this limitation works for only the limited words, first; afterwards, this effect will work for the other non-limited terms because this effect is expected to be propagated to the other non-limited terms whose meanings are similar to any of the limited words thorough the learning process. To realize this idea in a practical way, we utilize the Lexical Initialization (Section 2.3) in LEXIL.

2.3 LEXIL: Lexical Initialization Learning

This section describes the learning strategy of the SINN. Overall process is described in Algorithm 1.

Training First, to render the WCSL to represent the word-level contextual sentiment scores, the SINN is learned using the following $Lo^{\mathbf{Q}}$ as a loss function

$$Lo^{\mathbf{Q}} := SCE(\sum_{i=1}^{L}\sum_{t=1}^{T} c_{it}^{\mathbf{Q}}, d^{\mathbf{Q}}) \tag{7}$$

where $SCE(a, b)$ means the sigmoid cross-entropy between a and b. Through the learning with this $L^{\mathbf{Q}}$, the values in WCSL (= $\{\{c_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$) learn to represent word-level contextual sentiment scores (Proposition A.2).

Lexical Initialization Only the learning through $Lo^{\mathbf{Q}}$ cannot render the *WOSL*, *LCSL*, and *GCSL* to represent the corresponding scores. To solve this problem, we initialize the values in $\boldsymbol{W}^p$ using the prepared small word sentiment dictionary as follows (process 2 in Algorithm 1):

$$w_i^p \leftarrow \begin{cases} PS(w_i) & (w_i \in S^d) \\ 0 & (\text{otherwise}) \end{cases} \tag{8}$$

Algorithm 1 LEXIL: Lexical Initialization Learning

```
for $i \leftarrow 1$ to $v$ do
    Initialize $w_i^p$ as Eq (8);
Learn SINN using the gradient values by $Lo^{\mathbf{Q}}$;
```

where $PS(w_i)$ is the sentiment score for word w_i given by the word sentiment dictionary, and S^d is a set of words included in the dictionary.

Let $\Omega(S^d)$ be a set of word w_j that satisfies the $\min_{w_i \in S^d} \|\boldsymbol{w}_i^{em} - \boldsymbol{w}_j^{em}\|_2 < \delta$ where δ is sufficiently small. The lexical initialization is expected to improve the interpretability in LWCL, WOSL, and GWCL as follows.

A) LWCL By the effect of lexical initializaion, LWCL is expected to learn the sentiment shift for words in S^d and $\Omega(S^d)$ through LEXIL. (Propositions A.5 and A.6).

B) WOSL As a result, WOCL learns word-level original sentiment for words in $\Omega(S^d)$ through LEXIL, because the appropriate cases were decided for them (Proposition A.7).

C) GWCL GWCL learns to represent global word-level context through LEXIL because $\alpha_{it}^{\mathbf{Q}}$ is expected to become large in a case where $w_{it}^{\mathbf{Q}} \in \Omega(S^d)$ and any of the similar terms to $w_{it}^{\mathbf{Q}}$ has a strong sentiment (Proposition A.4). This manner is known to be natural for humans (Y. Zou 2018).

Through LEXIL, the number of words where WOSL, WLCSL, and GWCL can represent their corresponding sentiments (= $|\Omega(S^d)|$) becomes large gradually. After the learning has finished, we can extract word-level contextual sentiment scores from *WCSL* through extracting the word-level original sentiment, local word-level context, and global word-level context scores from WOSL, LWCL, and GWCL, as described in Appendix A.

3 Experimental Intepretability Evaluation

This section experimentally evaluates the proposed method in terms of the interpretability in A) WOSL, B) LWCL, and C) GWCL using real textual datasets.

3.1 Text Corpus

We used the following four textual corpora including reviews and their sentiment tags for evaluation.

1) EcoRev I and II. These two datasets are composed of comments on current (I) and future (II) economic trends and their positive or negative sentiment tags[1].

2) Yahoo reviews. This dataset is composed of comments on stocks and their long (positive) or short (negative) attitude tags, extracted from financial micro-blogs.[2]

3) Sentiment 140. This dataset contains tweets and their positive or negative sentiment tags.[3]

EcoRevs and Yahoo review are Japanese datasets, and Sentiment 140 is English. We used them to verify whether the SINN can be applied irrespective of the language or domain. We divided each dataset into training, validation, and test datasets, as outlined in Table 1.

3.2 Evaluation Metrics

After developing the SINN with the training and validation datasets, we evaluated the interpretability in A) WOSL, B) LWCL, and C) GWCL as follows.

A) Evaluation for WOSL For this evaluation, we used the economic, Yahoo, and LEX word polarity list[4], which include words along with their positive or negative polarities. The economic and Yahoo word polarity lists include Japanese economic terms, and the LEX word polarity list includes English terms. If we used the EcoRev I or II, Yahoo reviews, and Sentiment 140 in training, then, we utilized the economic, Yahoo, and LEX word polarity list, respectively. We used only the terms that appeared in the training dataset and not used in *LEXIL*. Table 1 summarizes the number of words used in this evaluation. We evaluated the interpretability of the WOSL based on the agreement between the polarities of word w_i (= answer) and w_i^p (= prediction) and used the macro F_1 score for the evaluation basis.

B) Evaluation for LWCL We prepared the Economy, Yahoo, and message annotated datasets for this evaluation. The Economy annotated dataset scontains 2,200 reviews (1,100 positive and 1,100 negative) from the test dataset of EcoRev I. The Yahoo annotated dataset includes 1,520 reviews (760 positive and 760 negative) from the test dataset of Yahoo reviews. The message annotated dataset has 10,258 reviews obtained from the test datasets in the SemEval tasks (Nakov et al. 2013; Rosenthal et al. 2014). In these datasets, part of the terms in the reviews had word-level sentiment shift tags indicating whether the sentiments of the terms were shifted (1: shifted) or not (0: non-shifted) as follows.

(1) In total, we are in a $bull^{(0)}$ market.

(2) This room is not $clean^{(1)}$.

(3) Products in this shop are too $expensive^{(1)}$.

Using these tags, we evaluated the interpretability of the LWCL according to the agreement between the sentiment shift tag of $w_{it}^{\mathbf{Q}}$ and the polarity of $s_{it}^{\mathbf{Q}}$ (shifted: $s_{it}^{\mathbf{Q}} < 0$ and non-shifted: $s_{it}^{\mathbf{Q}} > 0$). We used the macro F_1 score for the evaluation basis.

C) Evaluation for GWCL We used the global important point tags included in the Economy and Yahoo annotated datasets for this evaluation, which indicate whether each term in a review is important (1) or not (0) for deciding the document-level polarity of the review as follows.

(1) $We^{(0)}\ are^{(0)}\ in^{(0)}\ a^{(0)}\ bull^{(1)}\ market^{(1)}$.

(2) $This^{(0)}\ room^{(0)}\ is^{(0)}\ not^{(1)}\ clean^{(1)}$.

Using these tags, we evaluated the interpretability of the GWSL based on the correlation between $\{\alpha_t^{\mathbf{Q}}\}_{t=1}^n$ and the word-level global important points. We used the Pearson correlation for this evaluation.

In the evaluations for the LWCL and GWCL, we used the Economy, Yahoo, and message annotated datasets when we developed SINN with the EcoReviews, Yahoo reviews, and Sentiment 140, respectively. We only employed tags of terms that were not used in LEXIL and appeared in the training dataset. Table 1 summarizes the numbers of tags used.

[1] https://www5.cao.go.jp/keizai3/watcher-e/index-e.html

[2] http://textream.yahoo.co.jp

[3] https://www.kaggle.com/kazanova/sentiment140

[4] http://quanteda.io/reference/data_dictionary_LSD2015.html

3.3 SINN Development Setting

We developed the SINN using each training and validation datasets in the following settings.

LEXIL LEXIL used part of Japanese financial word sentiment dictionary (JFWS dict) and the Vader word sentiment dictionary (Vader dict) (Hutto and Gilbert 2014). These dictionaries contain words with sentiment scores. After excluding words with zero sentiment scores from these dictionaries, we extracted 200 words that appeared mostly in each training dataset from them and used their sentiment scores in LEXIL. The percentage of sentences covered by the above 200 terms was 3.4%, 4.1%, 0.7% and 7.5% in EcoRev I, EcoRev II, Yahoo, and Sentiment 140, respectively. To analyze the results when LEXIL used fewer words, we evaluated the SINNs developed with 50, 100, or 200 words in LEXIL: SINN (50), SINN (100) or SINN (200).

Others We calculated the word embedding matrix $\boldsymbol{W}^{em}$ with the skip-gram method (window size = 5) based on each textual corpus. We set the dimension of the hidden and embedding vectors to 200 and epoch to 50 with early stopping. We used the mean score of the five trials for evaluation.

3.4 Baseline

As explained in Section 2.2, the Lexical Initialization is expected to be important for realizing the interpretability of the SINN. To investigate its effect, we compared the results of the SINN with that of the following $SINN^{Base}$. The structure of $SINN^{Base}$ is the same as that of the SINN; however, it is different from SINN in that the values of $\boldsymbol{W}^{p}$ were initialized according to $U(-1, 1)$ where $U(a, b)$ is a uniform distribution between a and b, that is, $SINN^{Base}$ is developed without the Lexical Initialization.

A) Interpretability in WOSL To evaluate the interpretability of WOSL, we compared the results of the SINN and following word-level original sentiment analysis methods: PMI (Mohammad, Kiritchenko, and Zhu 2013), logistic fixed weight model (LFW) (Vo and Zhang 2016), sentiment-oriented NN (SONN) (Li 2017), and GINN (Ito et al. 2018). PMI is a statistical analysis method, while the others are interpretable NN based methods.

B) Interpretability in LWCL To evaluate the interpretability of LWCL, we compared the results of the SINN with that of the baseline and NegRNN methods. In the baseline, we predicted $w_{it}^{\mathbf{Q}}$ as "shifted" if the document-level sentiment tag of $\mathbf{Q}$ predicted by the RNN and sentiment tag of the word $w_{it}^{\mathbf{Q}}$ assigned by the PMI were different and as "not shifted" in other cases. In NegRNN, we first developed the polarity shifting training data using the weighed frequency odds method (Li et al. 2010), and then developed the RNN that predicts polarity shifts (Fancellu, Lopez, and Webber 2016), and used this for prediction.

C) Interpretability in GWCL To evaluate the interpretability of GWCL, we compared the evaluation result of SINN with that of the methods using the attention-based NNs: ATT (Yang et al. 2016), HN-ATT (Yang et al. 2016), SNNN (Hu et al. 2018), and LBSA (Y. Zou 2018). We used the attention score of each model as the global word-level context score.

Table 1: Dataset details for Text Corpus and Annotated data

Text Corpus	EcoRev I	EcoRev II	Yahoo	Sentiment 140
Training				
positive reviews	20,000	35,000	30,612	650,000
negative reviews	20,000	35,000	9,388	650,000
Validation				
positive reviews	2,000	2,000	3,387	50,000
negative reviews	2,000	2,000	1,613	50,000
Test				
positive reviews	4,000	4,000	7,538	100,000
negative reviews	4,000	4,000	2,462	100,000
vocabulary size v	8,071	11,130	33,080	71,316

Annotated data	EcoRev I	EcoRev II	Yahoo	Sentiment 140	
word polarity list					
Positive	348	337	422	1,843	
Negative	391	387	372	947	
sentiment shift tags					
Shifted tags	872	859	378	429	
Non-shifted tags	3,762	3,740	2,391	4,504	
word-level global important point tags					
Important tags (1)	6,632	6,631	1,526	-	
Unimportant tags (0)	62,652	62,652	48,890	-	
word-level and phrase-level contextual polarity tags					
Level	word	word	word	word	phrase
Shifted Negative	776	756	227	169	-
Non-shifted Negative	1,491	1483	1,187	1,294	-
Shifted Positive	96	96	151	260	-
Non-shifted Positive	2,271	2179	1,204	3,210	-
Negative (total)	2,267	2239	1,414	1,463	3,634
Positive (total)	2,367	2,275	1,355	3,470	5,907

3.5 Result and Discussion

Tables 2 indicate the results. The SINN significantly outperformed the other methods in most cases ($p < 0.05$ in five trials), demonstrating the high interpretability of the SINN. Moreover, the results of the SINNs and $SINN^{Base}$ demonstrate that the Lexical Initialization was important for realizing the interpretability of the SINN, as expected.

4 Experimental Evaluation for Word-level Contextual Sentiment Analysis Ability

This section experimentally evaluates the WCSA ability of the SINN in terms of the A) contextual word-level polarity, B) phrase-level polarity, and C) document-level polarity.

4.1 Evaluation Metrics

A) Contextual word-level polarity In the evaluation from this aspect, we evaluated the SINN in terms of the agreement between the polarity of word-level contextual sentiment for $w_{it}^{\mathbf{Q}}$ and the positive or negative polarity of $c_{it}^{\mathbf{Q}}$. We used the word-level contextual polarity tags included in the annotation datasets for this evaluation. They indicate the positive or negative word-level contextual polarities as follows.

(1) In total, we are in a $bull^{+}$ market.

(2) This room is not $clean^{-}$.

(3) Products in this shop are too $expensive^{-}$.

We used the macro average scores between the macro F1 score for the shifted terms and that for the non-shifted terms

Table 2: Evaluation Result for Interpretability

(A) Evaluation Result for WOSL

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
PMI	.734	.745	.793	.733
LFW	.715	.740	.766	.725
SONN	.702	.724	.725	.705
GINN	.723	755	754	.735
$SINN^{Base}$	.492	.513	.487	.444
SINN (200)	**.839**	**.856**	**.817**	**.737**
SINN (100)	**.842**	**.854**	**.816**	**.742**
SINN (50)	**.844**	**.854**	**.802**	**.751**

(B) Evaluation Result for LWCL

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
Baseline	.660	.712	.579	.560
NegRNN	.536	.626	.564	.558
$SINN^{Base}$	.350	.440	.495	.365
SINN (200)	**.800**	**.821**	**.646**	**.759**
SINN (100)	**.815**	**.857**	**.670**	**.742**
SINN (50)	**.776**	**.837**	**.659**	**.739**

(C) Evaluation Result for GWCL

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
ATT	-.015	-.081	.062	–
HN-ATT	.108	.188	.262	–
SNNN	.281	.456	.192	–
LBSA	.333	.344	**.405**	–
$SINN^{Base}$	.053	.131	.017	–
SINN (200)	**.588**	**.508**	.278	–
SINN (100)	**.637**	**.535**	.285	–
SINN (50)	**.602**	**.522**	.263	–

for the evaluation basis to test whether each method could accurately correspond to both shifted and non-shifted terms. We excluded the terms used in the Lexical Initialization, for fairness in comparison with the other methods.

B) Phrase-level polarity In the evaluation from this aspect, we evaluated the SINN in terms of the agreement between the polarity of the phrase-level sentiment for a phrase $\{w_{im}^{\mathbf{Q}}, \cdots, w_{in}^{\mathbf{Q}}\}$ and the polarity of $\sum_{t=n}^{m} c_{it}^{\mathbf{Q}}$ using the phrase-level polarity tags in the message annotated dataset. These tags indicate the positive or negative phrase-level polarity as follows.

(1) In total, we are in a $\{bull\ market\}^{+}$.

(2) This room is $\{not\ clean\}^{-}$.

(3) Products in this shop are $\{too\ expensive\}^{-}$.

We used the macro F1 score for the evaluation basis.

C) Document-level polarity In the evaluation from this aspect, we evaluated the SINN in terms of the agreement between the positive or negative polarity of the review **Q** and the polarity of $\sum_{i=1}^{L}\sum_{t=1}^{T} c_{it}^{\mathbf{Q}}$. We applied the document-level sentiment tags of reviews in test datasets for this evaluation and used the macro F_1 score as the evaluation basis.

Table 1 summarizes the numbers of tags used in the evaluations A), B), and C).

4.2 Comparison Methods

We compared the result of SINN with those from the following word-level sentiment analysis methods: PMI, LFW, SONN, GINN, Grad + RNN(Karen, Andrea, and Zisserman 2013), LRP + RNN (Arras et al. 2017), and IntGrad + RNN (Sundararajan, Taly, and Yan 2017), for this evaluation. The last three approaches are the developed LSTM interpretation-based approaches.

4.3 Result and Discussion

Tables 3 summarize the results. The SINN significantly outperformed the other methods ($p < 0.05$ in five trials). These results demonstrate the high WCSA ability of SINN, though its interpretability is high as demonstrated in Section 3.

Table 3: Evaluation Result for WCSA Ability

Evaluation Result in (A) Word-level polarity or (B) Phrase-level polarity

	EcoRev I	EcoRev II	Yahoo	Sentiment 140	
Level	word	word	word	word	phrase
PMI	.578	.548	.575	.631	.822
Grad + RNN	.578	.621	.601	.681	.743
IntGrad + RNN	.607	.621	.625	.679	.796
LRP + RNN	.597	.518	.579	.638	.808
LFW	.549	.545	.578	.587	.749
SONN	.555	.542	.566	.600	.787
GINN	.569	.555	.577	.623	.831
SINN (200)	**.719**	**.741**	**.651**	**.787**	**.863**
SINN (100)	**.724**	**.753**	**.649**	**.786**	**.847**
SINN (50)	**.695**	**.752**	**.640**	**.772**	**.855**

(C) Evaluation Result in Document-level polarity

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
LR	.878	.879	.741	.785
LFW	.876	.840	.751	.745
SONN	.863	.876	.717	.776
Grad + RNN	.870	.899	.724	.718
IntGrad + RNN	.909	.929	.750	.755
LRP + RNN	.909	.909	.751	.818
SINN (200)	**.928**	**.942**	**.766**	**.834**
SINN (100)	**.929**	**.941**	**.763**	**.832**
SINN (50)	**.928**	**.944**	**.761**	**.831**

4.4 Output Example

We experimentally demonstrate that both the interpretability and WCSA ability are high in SINN. We then introduce the text-visualization examples produced by SINN (Figs. 3). Like these examples, SINN can explain its WCSA results using the word-level original sentiment, local word-level contexts, and global word-level context scores.

From the first example in Japanese, we can see that the word-level contextual sentiment of "Fuel (Increase)" is strongly negative because its word-level original sentiment and local word-level context scores are positive and negative, respectively, and its global word-level context score is large. In addition, we can see that its sentiment shift occurs due to the left-oriented (i.e., backward) sentiment shift by "Nai (Not)" from the values in the LWCL. On the other hand, in the second example in English, we can see that the word-level contextual sentiment of "great" is strongly negative because its word-level original sentiment and local word-level context scores are positive and negative, respectively, and its global word-level context score is large. Moreover, we can see that its sentiment shift occurs due to the right-oriented (i.e., forward) sentiment shift by "Not" from the values in the LWCL.

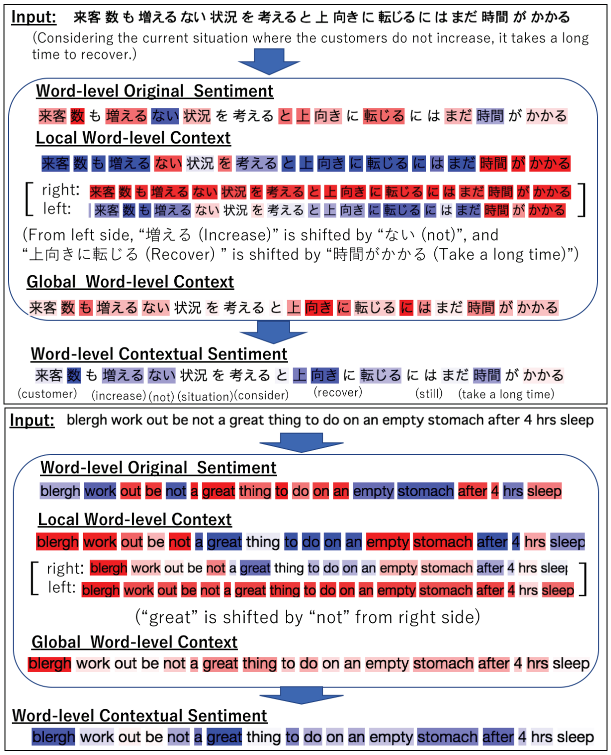
Input: 来客 数 も 増える ない 状況 を 考えると 上 向き に 転じる に は まだ 時間 が かかる
(Considering the current situation where the customers do not increase, it takes a long time to recover.)
Word-level Original Sentiment
来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
Local Word-level Context
来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
right: 来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
left: 来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
(From left side, "増える (Increase)" is shifted by "ない (not)", and "上向きに転じる (Recover) " is shifted by "時間がかかる (Take a long time)")
Global Word-level Context
来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
Word-level Contextual Sentiment
来客 数 も 増える ない 状況 を 考える と 上 向き に 転じる に は まだ 時間 が かかる
(customer) (increase) (not) (situation) (consider) (recover) (still) (take a long time)
Input: blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
Word-level Original Sentiment
blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
Local Word-level Context
blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
right: blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
left: blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
("great" is shifted by "not" from right side)
Global Word-level Context
blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep
Word-level Contextual Sentiment
blergh work out be not a great thing to do on an empty stomach after 4 hrs sleep

Figure 3: Text-visualization example by SINN. Colors mean their polarities (red: positive, blue: negative). The upper and below are reviews in EcoRev and Sentiment 140.

5 Related work

Many previous studies exist for the WCSA. Methods using annotated contextual sentiment tags (Mohammad, Kiritchenko, and Zhu 2013; Nakov et al. 2013; Rosenthal et al. 2014; Schulder et al. 2017) or specific knowledge (Li et al. 2013; Wilson, Wiebe, and Hoffman 2005; Kiritchenko and Mohammad 2016) were proposed. However, they are not practical because such annotated dataset or specific knowledge is typically not available for analyzing specialized documents. Some methods for analyzing sentiment shifts without thier specific knowledge have been proposed (Ikeda, Takamura, and Okumura 2008; Li et al. 2010; Xia et al. 2016); however, these methods need a big word polarity dictionary, and therefore not practical. In contrast to these methods, our method does not require any specific knowledge of sentiment shifts or big word polarity dictionary. Other approaches include statistical analysis based methods (Labille, Alfarhood, and Gauch 2016), methods using interpretable NNs (Li 2017; Ito et al. 2018; Vo and Zhang 2016; Y. Zou 2018), and interpretation techniques of prediction models (Karen, Andrea, and Zisserman 2013; Bach et al. 2017; Ribeiro, Singh, and Guestrin 2016; Shrikumar, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017). However, these alone cannot explain their analysis results using word-level original sentiment and local and global word-level contexts and lack the interpretability.

6 Conclusion

This study proposed a NN architecture called SINN. The SINN can extract word-level contextual sentiment through extracting word-level original sentiment and local and global word-level contexts; therefore, the SINN is interpretable. To realize the SINN practically, we proposed a novel learning strategy called LEXIL. We experimentally demonstrated that LEXIL was effective for improving the interpretability of SINN as well as that both the interpretability and WCSA ability of SINN were high. The SINN outperformed the comparative methods in the WCSA task on several domain datasets including Japanese and English datasets, while also featuring high interpretability. In the future, we will apply our SINN into other domain or language datasets. The dataset, code, and details will be available in http://bit.ly/SINN20190904.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Number JP17J04768.

A Theoretical Analysis in LEXIL

This section briefly describes theoretical analysis result in LEXIL. Before the explanation, we define several symbols. See the supplementary material for details and proofs. Let us define $R(\cdot)$, $PN(\cdot)$, and Condition A.1 as follows.

$$R(w_{it}^{\mathbf{Q}}) := \begin{cases} -1 & (sentiment\ of\ w_{it}^{\mathbf{Q}}\ is\ shifted) \\ 1 & (otherwise) \end{cases} .$$

$$PN(w_{it}^{\mathbf{Q}}) := \begin{cases} 1 & (\mathrm{sign}(d^{\mathbf{Q}} - 0.5) \neq R(w_{it}^{\mathbf{Q}})) \\ -1 & (\mathrm{sign}(d^{\mathbf{Q}} - 0.5) = R(w_{it}^{\mathbf{Q}})) \end{cases} .$$

Condition A.1 $w_i^p \begin{cases} > 0 & (OS(w_i^p) > 0) \\ < 0 & (OS(w_i^p) < 0) \end{cases}$ *is established where* $OS(w_j^p) := E[PN(w_{it}^{\mathbf{Q}}) | w_{it}^{\mathbf{Q}} = w_j^p, \mathbf{Q} \in \Omega^{tr}]$ *and* Ω^{tr} *is a set of reviews in a training dataset.*

Here, $PN(w_{it}^{\mathbf{Q}}) = 1$ denotes the case where the sentiment of $w_{it}^{\mathbf{Q}}$ is shifted in a negative review or the sentiment of $w_{it}^{\mathbf{Q}}$ is not shifted in a positive review, and $PN(w_{it}^{\mathbf{Q}}) = -1$ denotes the opposite case. Then, following three propositions are satisfied.

Proposition A.2 $\begin{cases} \frac{\partial Lo^{\mathbf{Q}}}{\partial c_{it}^{\mathbf{Q}}} < 0 & (d^{\mathbf{Q}} = 1) \\ \frac{\partial Lo^{\mathbf{Q}}}{\partial c_{it}^{\mathbf{Q}}} > 0 & (d^{\mathbf{Q}} = 0) \end{cases}$ *is satisfied.*

Proposition A.3 *If the Condition A.1 , is satisfied for every word* $w_i \in S^d$*, then, for every* $w_{it} \in \Omega(S^d)$*,*

$$\begin{cases} E[w_{I(w_{it})}^p] > 0 & (OS(w_{I(w_{it})}^p) > 0) \\ E[w_{I(w_{it})}^p] < 0 & (OS(w_{I(w_{it})}^p) < 0) \end{cases} \quad and \tag{9}$$

$$\begin{cases} E[s_{it}^{\mathbf{Q}}] > 0 & (R(w_{it}^{\mathbf{Q}}) > 0) \\ E[s_{it}^{\mathbf{Q}}] < 0 & (R(w_{it}^{\mathbf{Q}}) < 0) \end{cases} \tag{10}$$

are satisfied after sufficient iterations through LEXIL.

Proposition A.4 *After the sufficient iterations in LEXIL,* $E[\alpha_{it}^{\mathbf{Q}} | w_{it}^{\mathbf{Q}} \in \Omega^*(S^d)] > E[\alpha_{it}^{\mathbf{Q}} | w_{it}^{\mathbf{Q}} \notin \Omega^*(S^d)]$ *where* $\Omega^*(S^d)$ *is a subset of* $\Omega(S^d)$ *where if* $w_{it} \in \Omega^*(S^d)$*, then,*

$\max_{w_j \in \Theta(w_{it}^{\mathbf{Q}},\delta)} w_j^p > a$ *where* a *is sufficiently large and* $\Theta(w_{it}^{\mathbf{Q}},\delta)$ *is a set of words that satisfy* $\|\boldsymbol{e}_{it}^{\mathbf{Q}} - \boldsymbol{w}_j^{em}\|_2 < \delta$ *where* δ *is saficiently small, is satisfied.*

They indicate that WCSL, WOSL, LWCL, and GWCL learn to represent the corresponding scores in an ideal case. Moreover, this analysis suggests that the quality of the word sentiment dictionary is important for the success of propagation, where $|S^d|$ should not be too small and each word in S^d must satisfy Condition A.1. Proposition A.3 can be explained from the following propositions. See the supplementary material for the details.

Proposition A.5 *If Condition A.1 is satisfied for word* $w_{it}^{\mathbf{Q}}$*, then, Eq (10) is satisfied for* $w_{it}^{\mathbf{Q}}$*.*

Proposition A.6 *If* w_i *satisfies Condition A.1 and Eq (10), then Eq (10) is satisfied for* $w_j \in \Theta(w_{it}^{\mathbf{Q}},\delta)$ *where* δ *is sufficiently small.*

Proposition A.7 *If Eq (10) is satisfied for* w_i*, then, Eq (9) is satisfied for* w_i *and becomes to satisfy Condition A.1.*

References

Arras, L.; Montavon, G.; Muller, K. R.; and Samek, W. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *EMNLP Workshop*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Muller, K. R.; and Samek, W. 2017. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PLOS ONE* 10(7):1–46.

Fancellu, F.; Lopez, A.; and Webber, B. 2016. Neural networks for negation scope detection. In *ACL 2016*.

Hu, Q.; Zhou, J.; Chen, Q.; and He, L. 2018. Snnn: Promoting word sentiment and negation in neural sentiment classification. In *AAAI 2018*.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM-14*.

Ikeda, D.; Takamura, H.; and Okumura, M. 2008. Learning to shift the polarity of words for sentiment classification. In *IJCNLP 2008*, 50–57.

Ito, T.; Sakaji, H.; Tsubouchi, K.; Izumi, K.; and Yamashita, T. 2018. Text-visualizing neural network model: Understanding online financial textual data. In *PAKDD 2018*.

Karen, S.; Andrea, V.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*.

Kiritchenko, S., and Mohammad, S. M. 2016. The effect of negators, modals, and degree adverbs on sentiment composition. In *NAACL-HLT 2016*, 43–52.

Labille, K.; Alfarhood, S.; and Gauch, S. 2016. Estimating sentiment via probability and information theory. In *KDIR 2016*, 121–129.

Li, S.; Yat, S.; Lee, M.; Chen, Y.; Huang, C. R.; and Wang, G. 2010. Sentiment classification and polarity shifting. In *COLING 2010*, 635–643.

Li, S.; Wang, Z.; Lee, S. Y. M.; and Huang, C.-R. 2013. Sentiment classification with polarity shifting detection. In *IALP 2013*, 129–132.

Li, Q. 2017. Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *CoNLL 2017*, 301–310.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*.

Mohammad, S.; Kiritchenko, S.; and Zhu, X. D. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval-2013*.

Nakov, P.; Rosenthal, S.; Kozareva, .; Stoyanov, V.; Ritter, A.; and Wilson, T. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *SemEval 2013*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ”why should i trust you?” explaining the predictions of any classifier. In *KDD*.

Rosenthal, S.; Nakov, P.; Ritter, A.; and Stoyanov, V. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *SemEval 2014*.

Schulder, M.; Wiegand, M.; Ruppenhofer, J.; and Roth, B. 2017. Towards bootstrapping a polarity shifter lexicon using linguistic features. In *IJCNLP 2017*, 624–633.

Schuster, M., and Paliwal, K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *ICML*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for simplicity: The all convolutional net. In *ICLR Workshop*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *ICML*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2016. Attention is all you need. In *NIPS 2017*.

Vo, D. T., and Zhang, Y. 2016. Don't count, predict! an automatic approach to learning sentiment lexicons for short text. In *ACL 2016*, 219–224.

Wilson, T.; Wiebe, J.; and Hoffman, P. 2005. Recognizing contextual polarity in phrase level sentiment analysis. In *EMNLP 2005*, 347–354.

Xia, R.; Xu, F.; Yu, J.; Qi, Y.; and Cambria, E. 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing and Management* 52(1):36–45.

Y. Zou, T. Gui, Q. Z. X. H. 2018. A lexicon-based supervised attention model for neural sentiment analysis. In *COLING 2018*.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL 2016*.