CSNN: Contextual Sentiment Neural Network

Tomoki Ito*, Kota Tsubouchi†, Hiroki Sakaji*, Kiyoshi Izumi* and Tatsuo Yamashita†
*Graduate School of Engineering,The University of Tokyo and †Yahoo Japan Corporation
Email: m2015titoh@socsim.org

***Abstract*—Although deep neural networks are excellent for text sentiment analysis, their applications in real-world practice are occasionally limited owing to their black-box property. In response, we propose a novel neural network model called contextual sentiment neural network (CSNN) model that can explain the process of its sentiment analysis prediction in a way that humans find natural and agreeable. The CSNN has the following interpretable layers: the word-level original sentiment layer, word-level sentiment shift layer, word-level local contextual sentiment layer, word-level global importance layer, and word-level global contextual sentiment layer. Because of these layers, this network can explain the process of its document-level sentiment analysis results in a human-like way using these layers. Realizing the interpretability of each layer in the CSNN is a crucial problem in the development of this CSNN because the general back-propagation method cannot realize such interpretability. To realize this interpretability, we propose a novel learning strategy called initialization propagation (IP) learning. Using real textual datasets, we experimentally demonstrate that the proposed IP learning is effective for improving the interpretability of each layer in CSNN. We then experimentally demonstrate that both the predictability and explanation ability of the CSNN are high.**

***Index Terms*—Interpretable Neural Networks, Text-mining, Support System**

I. Introduction

Massive web documents such as micro-blogs and customer reviews are useful for public opinion sensing and trend analysis. The sentiment analysis approach (i.e., to automatically predict whether a review is overall positive or negative) has been commonly used in this area. Deep neural networks (DNNs) are some of the best-performing machine learning methods [1]. However, DNNs are often avoided in cases where explanations are required because these networks are generally considered as black-boxes. Thus, developing a high predictable neural network (NN) model that can explain the process of its prediction process in a human-like way is a critical problem. In the development of such NN model, we should consider how humans usually judge the positive or negative polarity of each review. As described in some previous linguistic researches [2], [3], it is well known that humans judge the positive or negative polarity of each review by extracting four types of word-level sentiment scores in the following order.

1) Word-level original sentiment: The sentiment that each word in a document originally has (e.g., scores in a word sentiment dictionary [4]).

2) Word-level local contextual sentiment: The positive or negative sentiment score of each term in a document after considering its sentiment shift, such as ”good” in ”not good” and ”goodness” in ”decrease the goodness.”

3) Word-level global contextual sentiment: The positive or negative sentiment score of each term after considering what part is important in the entire document (i.e., the global important point) and its sentiment shift, and

4) Document-level sentiment: The prediction results for positive or negative sentiment tags of reviews.

Therefore, to explain the prediction results in a form that humans feel natural and agreeable, we need to use the above four types of sentiments as shown in Fig. 1:

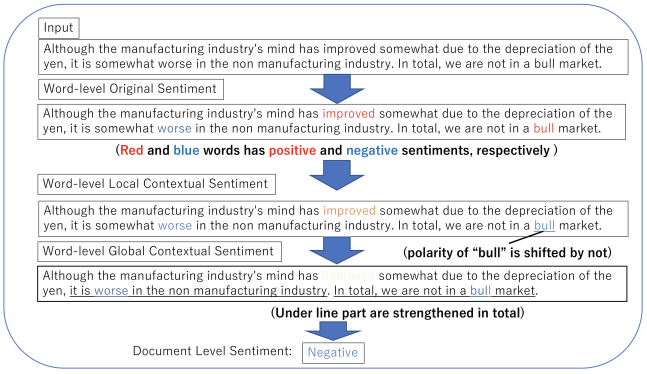

Fig. 1. Goal: development of neural network (NN) that can explain its prediction results using four types of sentiments

However, a method for developing NNs that can explain its predictions using these four types of sentiments is yet to be established. Many studies have been done to address the black-box property of the NNs [5]–[14]; however, it is hard to say that they have the sufficiently high explanation ability because they alone cannot describe these four types of sentiments.

To solve this problem, we propose a novel NN model called contextual sentiment NN (CSNN) and a novel learning strategy called initialization and propagation (IP) learning.

CSNN has the following four interpretable layers: word-level original sentiment layer (WOSL), sentiment shift layer (SSL), word-level local contextual word-level sentiment layer (WLCSL), global important point layer (GIL), and word-level global contextual sentiment layer (WGCSL) as shown in Fig. 2. The WOSL, WLCSL, and WGCSL represent the word-level original, local contextual, and global contextual sentiments of each term in a review, respectively. The SSL indicates whether a sentiment of each term in a review is shifted or not, and GIL indicates the global important points in a review. WOSL is represented in a dictionary manner. SSL and GIL are represented using long short-term memories (LSTM) cells [15]. The values

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00135

of WLCSL and WGCSL are represented by multiplying the values of WOSL and SSL, and by multiplying the values of WLCSL and GIL, respectively. Therefore, using these leyers,

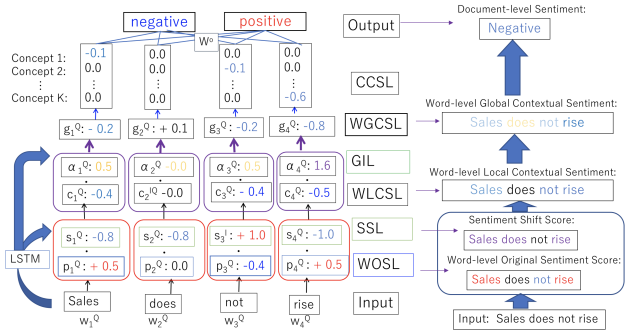

Fig. 2. CSNN

the CSNN can explain the process of the sentiment analysis prediction in a human agreeable form.

In developing this CSNN, realizing such interpretability for each layer is crucial. General back-propagation based learning using reviews and their sentiment tags cannot render each layer to represent corresponding sentiment. To realize this interpretability, we propose a novel learning strategy called initialization and propagation (IP) learning. IP learning requires only reviews, their sentiment tags, and a small word sentiment dictionary, and does not require any information for context. This is a valuable point in our approach.

The contributions of this paper are as summarized follows:

(1) We proposed a novel NN architecture called CSNN that can explain its sentiment analysis process in a form that humans find natural and agreeable,

(2) To realize the interpretability of CSNN, we proposed a novel learning strategy called IP learning, and

(3) We experimentally demonstrated that a) IP learning improved the explainability of the CSNN, and that (b) both the interpretability and predictability of the CSNN were high.

II. CSNN

This section introduce the proposed CSNN. A CSNN can be developed through IP learning (Section II-A) using a training dataset $\{(\mathbf{Q}_i, d^{\mathbf{Q}_i})\}_{i=1}^N$, and a small word sentiment dictionary. Note that N is the training data size, $\mathbf{Q}_i = \{w_t^{\mathbf{Q}_i}\}_{t=1}^n$ is a review, $d^{\mathbf{Q}_i}$ is its sentiment tag (1 is positive and 0 is negative), and $w_t^{\mathbf{Q}_i}$ is a tth term in $\mathbf{Q}_i$.

A. Structure of the CSNN

This section introduces the structure of the CSNN, which includes the WOSL, SSL, WLCSL, GIL, WGCSL, concept-level contextual sentiment layer (CCSL).

Notation. We first define several symbols. Let $\{w_i\}_{i=1}^v$ represent the terms that appear in a text corpus, and v be the vocabulary size. We define the vocabulary index of word w_i as $I(w_i)$. Therefore, $I(w_i) = i$. Let $\boldsymbol{w}_i^{em} \in \mathbb{R}^e$ be an embedding representation of word w_i, and the embedding matrix $\boldsymbol{W}^{em} \in \mathbb{R}^{v\times e}$ be $[\boldsymbol{w}_1^{emT}, \cdots, \boldsymbol{w}_v^{emT}]^T$. where e is the dimension size of word embedding and $\|\boldsymbol{w}_i^{em}\|_2 = 1$. $\boldsymbol{W}^{em}$ is the constant value obtained using the skip-gram method [16] and each text corpus in a training dataset.

1) WOSL: This layer represents word-level original sentiment representations $\{p_t^{\mathbf{Q}}\}_{t=1}^n$ as

$$p_t^{\mathbf{Q}} = w^p_{I(w_t^{\mathbf{Q}})}$$

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

2) SSL: This layer represents the word-level sentiment shift scores $\{s_t^{\mathbf{Q}}\}_{t=1}^n$ ($s_t^{\mathbf{Q}} < 0$: shifted, and $s_t^{\mathbf{Q}} > 0$: not shifted) as

$$\overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}} = \overrightarrow{\mathrm{LSTM}}(\boldsymbol{e}_t^{\mathbf{Q}}), \overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}} = \overleftarrow{\mathrm{LSTM}}(\boldsymbol{e}_t^{\mathbf{Q}}),$$
$$\overleftarrow{s}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{left} \cdot \overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}}), \overrightarrow{s}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{right} \cdot \overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}}),$$
$$s_t^{\mathbf{Q}} := \overrightarrow{s}_t^{\mathbf{Q}} \cdot \overleftarrow{s}_t^{\mathbf{Q}}.$$

Here, $\boldsymbol{e}_t^{\mathbf{Q}}$ represents the embedding of $w_t^{\mathbf{Q}}$ from $\boldsymbol{W}^{em}$, $\overrightarrow{\mathrm{LSTM}}$ and $\overleftarrow{\mathrm{LSTM}}$ represents the conversions by forward and backward LSTMs, $\boldsymbol{v}^{right}, \boldsymbol{v}^{left} \in \mathbb{R}^e$ are parameter values. $\overrightarrow{s}_t^{\mathbf{Q}}$ and $\overleftarrow{s}_t^{\mathbf{Q}}$ denote whether or not the sentiment of $w_t^{\mathbf{Q}}$ is shifted by the left-side and right-side terms of $w_t^{\mathbf{Q}}$: $\{w_{t'}^{\mathbf{Q}}\}_{t'=1}^{t-1}$ and $\{w_{t'}^{\mathbf{Q}}\}_{t'=t+1}^n$, respectively.

3) GIL: This layer represents the word-level global important point representations $\{\alpha_t^{\mathbf{Q}}\}_{t=1}^n$ using a revised self-attention mechanism [17], [18] as

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

4) WLCSL: This layer represents word-level local contextual sentiment representations $\{c_t^{\mathbf{Q}}\}_{t=1}^n$ as

$$c_t^{\mathbf{Q}} = s_t^{\mathbf{Q}} \cdot p_t^{\mathbf{Q}}.$$

5) WGCSL: This layer represents word-level global contextual sentiment representations $\{g_t^{\mathbf{Q}}\}_{t=1}^n$ as

$$g_t^{\mathbf{Q}} := c_t^{\mathbf{Q}} \alpha_t^{\mathbf{Q}}.$$

6) CCSL: This layer represents the contextual concept-level sentiment representation $\boldsymbol{v}^{\mathbf{Q}} := \sum_{t=1}^n g_t^{\mathbf{Q}} \boldsymbol{b}_t^{\mathbf{Q}}$ where $\boldsymbol{b}_t^{\mathbf{Q}} := \max(\mathrm{Softmax}(\boldsymbol{W}_c \boldsymbol{e}_t^{\mathbf{Q}} - t_c), 0)$, $\boldsymbol{v}_t^{\mathbf{Q}} \in \mathbb{R}^K$, $\boldsymbol{b}_t^{\mathbf{Q}} \in \mathbb{R}^K$, $t_c > 0$ is a hyper-parameter value, $\boldsymbol{W}_c \in \mathbb{R}^{K\times e}$ is centroid vectors of $\{\boldsymbol{w}_i^{em}\}_{i=1}^v$ calculated using a spherical k-means method [19] where the cluster number is K.

7) Output: Then, CSNN outputs a predicted sentiment tag $y^{\mathbf{Q}} \in \{0(\mathrm{negative}), 1(\mathrm{positive})\}$ as

$$\boldsymbol{a}^{\mathbf{Q}} = \mathrm{Softmax}(\boldsymbol{W}^O \tanh(\boldsymbol{v}^{\mathbf{Q}})), y^{\mathbf{Q}} = \mathrm{argmax}\ \boldsymbol{a}^{\mathbf{Q}}$$

where $\boldsymbol{W}^O \in \mathbb{R}^{2\times K}$ is the parameter value.

B. Key Idea in IP learning

In developing CSNN, the realization of the interpretability in WOSL and SSL is especially difficult. Through the learning with $L^{\mathbf{Q}}$ and Update (will be defined later), WLCSL and WGCSL learn to represent corresponding sentiments. However, this learning strategy alone cannot realize the interpretability in WOSL and SSL because in the case where the

polarity of $c_t^{\mathbf{Q}}$ is accurately negative, the following two cases are possible: (1) $p_t^{\mathbf{Q}} > 0$ and $s_t^{\mathbf{Q}} < 0$, or (2) $p_t^{\mathbf{Q}} < 0$ and $s_t^{\mathbf{Q}} > 0$, and the accurate case cannot be chosen automatically in general learning. We assume that this problem can be solved by initially limiting the polarity of $p_t^{\mathbf{Q}}$ to the accurate case for a few words because this limitation leads to the accurate choice from the above two cases. Therefore, this limitation can lead to the learning of $s_t^{\mathbf{Q}}$ within the appropriate case. The effect of this limitation works for only the limited words, first; however, this effect is assumed to be propagated to the other non-limited terms whose meanings are similar to any of the limited words thorough learning, afterward. To realize this idea, we utilize the Init (will be defined later) in IP learning.

C. Initialization and Propagation (IP) Learning

This section describes the learning strategy of the SINN. Overall process is described in Algorithm 1 where $w_{i,j}^o$ is the (i,j) element of $\boldsymbol{W}^O$, and $L^{\mathbf{Q}}$ is the cross entropy between $\boldsymbol{a}^{\mathbf{Q}}$ and $d^{\mathbf{Q}}$. IP learning utilizes the two specific techniques called *Update* and *Init*. *Update* is a strategy for improving the interpretability in $WLCSL$ and $WGCSL$. *Init* is a strategy for improving the interpretability in $WOSL$ and GIL. Using both the *Update* and *Init*, the interpretability in SSL is also expected to be improved (as theoretically analyzed in Appendix A in the supplementary material).

1) Update: First, $\boldsymbol{W}^O$ is updated according to processes 6–7 in Algorithm 1. This makes $WLCSL$ and $WGCSL$ to represent the corresponding sentiment scores (Proposition A.3 in Appendix) without violating the learning process after sufficient iterations (Proposition A.7 in Appendix A).

2) Init: Then, $\boldsymbol{W}^p$ is initialized as process 2 in Algorithm 1, 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 from the dictionary. Init makes $WOSL$ and SSL represent the corresponding scores in the condition that Update is utilized.

Through this IP learning, for every word sufficiently similar to any of the words in S^d, the *WOSL*, *SSL*, *WLCSL*, *GIL*, and *WGCSL* learn to represent the corresponding scores, as theoretically analyzed in Appendix A. After the learning, the CSNN can explain its prediction result using these layers.

Algorithm 1 Initialization and Propagation (IP) Learning

1: **for** $i \leftarrow 1$ to v **do**
2: $\quad w_i^p \leftarrow \begin{cases} PS(w_i) & (w_i \in S^d) \\ 0 & (\text{otherwise}) \end{cases}$;
3: **while** learning has not been finished **do**
4: $\quad$ Update $\boldsymbol{W}^p$, $\boldsymbol{v}^{right}$, $\boldsymbol{v}^{left}$, $\boldsymbol{W}^O$ and the LSTM cells in CSNN using the gradient values by $L^{\mathbf{Q}}$. ;
5: $\quad$ **for** $k \leftarrow 1$ to K **do**
6: $\quad\quad$ **if** $w_{1,k}^o > 0$ **then** $w_{1,k}^o \leftarrow 0$;
7: $\quad\quad$ **if** $w_{2,k}^o < 0$ **then** $w_{2,k}^o \leftarrow 0$;

III. Experimental Evaluation

This section experimentally tests the explanation ability and predictability of the CSNN and investigate the effect of IP learning for the interpretability of the layers in the CSNN.

A. Dataset

1) Text Corpus: We used the following four textual corpora, including reviews and their sentiment tags, for this evaluation. They were used for developing CSNN.

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

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

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

EcoRevs and Yahoo review were Japanese datasets, and Sentiment 140 was an English dataset. We used them to verify whether the CSNN can be used irrespective of the language or domain. We divided each dataset into the training, validation, and test datasets, as presented in Table I.

2) Annotated Dataset: For this evaluation, we prepared the Economy, Yahoo, and message annotated datasets. The Economy annotated dataset has 2,200 reviews (1,100 positive and 1,100 negative) in the test dataset of EcoReviews I. The Yahoo annotated dataset has 1520 reviews (760 positive and 760 negative) in the test dataset of Yahoo reviews. The message annotated dataset has 10258 reviews obtained from the test datasets in SemEval tasks [20], [21]. These catasets included the word-level or phrase-level contextual sentiment tags, word-level sentiment shift tags, and word-level global important point tags. Word-level or phrase-level contextual sentiment tags indicate whether the word-level or phrase-level contextual sentiments of terms are positive or negative. Word-level sentiment shift tags indicate whether the sentiments of terms were shifted (1: shifted tags) or not (0: non-shifted tags). Word-level global important point tags indicates whether each term in a review is important (1) or not (0) for deciding the overall positive or negative polarity of the review. See the supplementary material for details.

B. CSNN Development Setting

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

Setting in Init. *Init* used a part of a Japanese financial word sentiment dictionary (JFWS dict) developed by six financial professionals and the Vader word sentiment dictionary (Vader dict) [4]. These dictionaries contain words and their sentiment scores. After we excluded the words with zero sentiment scores and those with absolute sentiment scores of less than 1.0 from JFWS dict and the Vader dict, respectively, we extracted most frequen 200 words in each training dataset from these dictionaries and used their sentiment scores in *Init*. To analyze the results in the cases where *Init* used fewer words, we evaluated the results with CSNNs developed with only 50 or 100, or 200 words: CSNN (50), CSNN (100) and CSNN (200).

[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

Other settings. We calculated the word embedding matrix $\boldsymbol{W}^{em}$ by the skip-gram method (window size = 5) [16] based on each textual dataset. We set the dimensions of the hidden and embedding vectors to 200, epoch to 50 with early stopping, K to $[100, 500, 1000]$, t_c to $1/K$. We determined the hyper-parameters using the validation data. We used the mean score of the five trials for the evaluations in this paper.

TABLE I
DATASET ORGANIZATION

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

C. Evaluation Metrics in Explanation ability

Evaluation Metric. We evaluated the explanation ability of the CSNN based on the validity in WOSL, SSL, WLCSL, GIL, and WGCSL in the following way.

1) WOSL: We evaluated the validity of WOSL based on the agreement between the polarities of word w_i and w_i^p using the economic, Yahoo, and LEX word polarity list[4]. These lists include words and their positive or negative polarities. LEX word-polarity list includes English terms, and the others include Japanese economic terms.

2) SSL: Using the sentiment shift tags, we evaluated the validity of the SSL based on the agreement between the sentiment shift tag of $w_t^{\mathbf{Q}}$ and the polarity of $s_t^{\mathbf{Q}} > 0$ (shifted: $w_i^p < 0$ and non-shifted: $w_i^p > 0$).

3) WLCSL: Using the word or phrase level contextual sentiment tags, we evaluated the validity of the WLCSL based on the agreement between the polarity of $c_t^{\mathbf{Q}}$ and the contextual word-level sentiment tag of $w_t^{\mathbf{Q}}$ or the agreement between the polarity of the summed scores for terms involved in each phrase accurately and its phrase-level sentiment. We used the micro and macro average scores between the macro F_1 score for shifted terms and that for non-shifted terms for the evaluation basis. We used the micro-average score to test whether each method could work in real situations, and macro-average score to test whether each method could accurately correspond to both shifted and non-shifted terms.

4) GIL: Using the gold word-level global important points, we evaluated the validity of GIL based on the correlation between $\{\alpha_t^{\mathbf{Q}}\}_{t=1}^n$ and gold word-level global important points. We used the Pearson correlation for this evaluation.

5) WGCSL: We evaluated the explanation validity of the WGCSL based on the agreement between the polarities of $\sum_{t=1}^n g_t^{\mathbf{Q}}$ and the document-level sentiment tag of $\mathbf{Q}$. We used the macro F_1 score as the evaluation basis.

In the above evaluations, we used the Economy, Yahoo, and message annotated datasets when developing CSNNs with the corresponding text corpus, respectively. We only employed tags of terms that were not used in Init and appeared in the training dataset. Table I summarizes the numbers of tags used.

Baselines. To evaluate the effect of IP learning, we compared the results of the CSNNs and those of the following baselines: $CSNN^{Base}$, $CSNN^{Rand}$, and $CSNN^{NoUp}$. The structures of these models are the same as that of CSNN; the differences are summarized as below.

I) $CSNN^{Base}$ is developed using the general backpropagation and without Update or Init strategy.

II) $CSNN^{Rand}$ is developed with only Update strategy.

III) $CSNN^{NoUp}$ is developed with only Init strategy.

Comparison Method. To evaluate the explanation ability of CSNN, we compared the evaluation result of CSNN with other comparative methods in each layer validity.

1) WOSL: This evaluation compared the CSNN with the other word-level original sentiment assignment methods, namely, PMI [22], logistic fixed weight model (LFW) [7], sentiment-oriented NN (SONN) [8], and gradient interpretable neural network (GINN) [9].

2) SSL: This evaluation compared the CSNN with the baseline and NegRNN methods. In the baseline, we predicted $w_t^{\mathbf{Q}}$ as "shifted" if the sentiment of $d^{\mathbf{Q}}$ predicted by the RNN and sentiment tag of $w_t^{\mathbf{Q}}$ assigned by the PMI were different and as "not shifted" in other cases. In NegRNN, we used the RNN that predicts polarity shifts [23] developed with the the polarity shifting training data created by the weighed frequency odds method [24].

3) WLCSL: This evaluation compared the CSNN with the other word-level sentiment assignment methods: PMI, LFW, SONN, GINN, Grad + a bidirectional LSTM model (RNN) [12], LRP + RNN [25], and IntGrad + RNN [11].

4) GIL: This evaluation compared the CSNN with the other word-level important point assignment methods using the RNNs using attention mechanism: word attention network (ATT) [26], hierarchical attention network (HN-ATT) [26], sentiment and negation neural network (SNNN) [27], and lexicon-based supervised attention (LBSA) [5]. SNNN and LBSA are set up in a form that the attention weights of terms

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

with the strong word-level original sentiment are strengthened. We used the attention score of each model as the score.

5) WGCSL: This evaluation compared the CSNN with the comparative methods used in the evaluation in WLCSL.

D. Evaluation Metrics in Predictability

Evaluation Metric. We evaluates the predictability of the CSNN based on whether it can predict the sentiment tags of reviews in each test dataset.

Comparison Method. We compared the CSNN and the following methods: logistic regression (LR), LFW [7], SONN [8], GINN [9], a bi-LSTM based RNN (RNN), convolutional NN (CNN) [1], ATT [26], HN-ATT [26], SNNN [27], LBSA [5]. We used the macro F_1 score as the evaluation basis.

E. Result and Discussion

1) Explanation ability and Predictability: Tables II summarize the results for explanation ability, indicating that the proposed CSNN outperformed the other methods in most cases. Table III summarizes the results, indicating that HN-ATT had greater predictability than the proposed CSNNs; however, CSNN (200) had greater predictability than LR and some deep NNs such as CNN and SNNN, and had predictability equivalent to that of ATT or LBSA.

These results demonstrate that the proposed CSNN has both the high explanation ability and high predictability.

2) Effect of IP learning: The results of CSNNs, $CSNN^{Base}$, $CSNN^{NoUp}$, and $CSNN^{Rand}$ for explainability demonstrate the effect of IP learning as follows. The $CSNN^{Rand}$ outperformed the $CSNN^{Base}$ in WLCSL and WGCSL, indicating that Update promoted the validity in WLCSL and WGCSL; whereas, the $CSNN^{NoUp}$ outperformed the $CSNN^{Base}$ in WOSL and GIL, indicating that Init promoted the validity in WOSL and GIL. Consequently, the validity in all the five layers were improved by using both Update and Init, and the CSNNs outperformed the $CSNN^{Base}$ in all the cases. This is the expected result as described in Section II-C (and Appendix A in the supplementary).

F. Text-Visualization Example

This section introduces some examples of text-visualization produced by the CSNN. Fig. 3 shows the text-visualization examples. Users can explain the CSNN's prediction process based on this type of text-visualizations.

IV. Related work

There are many studies for addressing the black-box property of the deep NNs. As a useful technique for explaining the prediction results of NNs, we can present methods for interpreting prediction models [10]–[13], [28], [29]. These methods calculated the gradient score of each input feature in the prediction and visualized an important feature in their predictions. The LRP method is one of the state-of-the-art methods. Interpretable NNs [5]–[8] are also useful in these aspects. However, the previous methods do not satisfy our purpose because they alone cannot represent all the word-level original, local contextual, and global contextual sentiments.

TABLE II
Evaluation Result for Explanation Ability

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
Evaluation Result of WOSL (Macro F_1 score)				
PMI	0.734	0.745	0.793	0.733
LFW	0.715	0.740	0.766	0.725
SONN	0.702	0.724	0.725	0.705
GINN	0.723	0.755	0.754	0.735
$CSNN^{Base}$	0.417	0.381	0.499	0.373
$CSNN^{NoUp}$	**0.832**	**0.846**	**0.798**	**0.754**
$CSNN^{Rand}$	0.452	0.543	0.460	0.430
CSNN (200)	**0.837**	**0.865**	**0.825**	**0.742**
CSNN (100)	**0.838**	**0.851**	**0.817**	**0.744**
CSNN (50)	**0.843**	**0.865**	**0.805**	**0.743**
Evaluation Result of SSL (Macro F_1 score)				
Baseline	0.660	0.712	0.579	0.560
NegRNN	0.536	0.626	0.564	0.558
$CSNN^{Base}$	0.661	0.311	0.244	0.314
$CSNN^{NoUp}$	0.374	0.246	0.360	0.417
$CSNN^{Rand}$	0.263	0.531	0.315	0.293
CSNN (200)	**0.777**	**0.804**	**0.691**	**0.743**
CSNN (100)	**0.780**	**0.816**	**0.681**	**0.751**
CSNN (50)	**0.784**	**0.809**	**0.675**	**0.762**

Evaluation Result of WLCSL (Macro F_1 score)

Level	EcoRev I word		EcoRev II word		Yahoo word		Sentiment 140 word		phrase
	macro	micro	macro	micro	macro	micro	micro	macro	
PMI	.792	.578	.788	.548	**.823**	.575	.854	.631	.822
Grad + RNN	.703	.578	.743	.621	.713	.601	79.3	68.1	74.3
IntGrad + RNN	.801	.607	.775	.621	.752	.625	.842	.679	.79.6
LRP + RNN	.805	.597	.741	.518	.761	.579	.834	.638	.808
LFW	.789	.549	.791	.545	.811	.578	.832	.587	.749
SONN	.767	.555	.788	.542	.769	.566	.866	.600	.787
GINN	.796	.569	.790	.555	.770	.577	.861	.623	.831
$CSNN^{Base}$	.378	.355	.626	.521	.522	.490	.612	.575	.595
$CSNN^{NoUp}$	.427	.416	.273	.316	.566	.526	.505	.509	.512
$CSNN^{Rand}$	*.714*	*.606*	*.763*	*.621*	*.674*	*.516*	*.810*	*.794*	*.748*
CSNN (200)	**.855**	**.676**	**.878**	**.711**	**.817**	**.669**	**.891**	**.788**	**.858**
CSNN (100)	**.849**	**.679**	**.879**	**.723**	**.812**	**.675**	**.893**	**.784**	**.862**
CSNN (50)	**.861**	**.692**	**.880**	**.719**	**.797**	**.670**	**.889**	**.788**	**.857**

Evaluation Result of GIL (Pearson Correlation)				
ATT	-0.015	-0.081	0.062	-
HN-ATT	0.108	0.188	0.262	-
SNNN	0.281	0.456	0.192	-
LBSA	0.333	0.344	**0.405**	-
$CSNN^{Base}$	0.014	0.170	0.171	-
$CSNN^{NoUp}$	**0.607**	**0.590**	**0.329**	-
$CSNN^{Rand}$	0.207	0.224	0.164	-
CSNN (200)	**0.595**	**0.580**	**0.325**	-
CSNN (100)	**0.584**	**0.567**	**0.308**	
CSNN (50)	**0.585**	**0.562**	**0.321**	-
Evaluation Result of WGCSL (Macro F_1 score)				
PMI	0.827	0.800	0.673	0.759
LFW	0.876	0.840	0.751	0.745
SONN	0.863	0.876	0.717	0.776
GINN	0.860	0.859	0.740	0.782
Grad + RNN	0.870	0.899	0.724	0.718
IntGrad + RNN	0.909	0.929	0.750	0.755
LRP + RNN	0.909	0.909	0.751	0.818
$CSNN^{Base}$	0.248	0.709	0.534	0.615
$CSNN^{NoUp}$	0.417	0.239	0.533	0.565
$CSNN^{Rand}$	**0.911**	**0.916**	**0.717**	**0.831**
CSNN (200)	**0.923**	**0.937**	**0.771**	**0.830**
CSNN (100)	**0.916**	**0.935**	**0.768**	**0.829**
CSNN (50)	**0.918**	**0.938**	**0.766**	**0.831**

V. Conclusion

This paper proposed a novel NN architecture called CSNN that can explain its prediction process. To realize the explainability of CSNN, we proposed a novel learning strategy called

TABLE III
F_1 SCORE RESULTS FOR THE PREDICTABILITY EVALUATION

	EcoRev I	EcoRev II	Yahoo	Sentiment 140
LR	0.878	0.879	0.741	0.785
CNN	0.894	0.911	0.757	0.820
RNN	0.922	0.932	0.749	**0.837**
ATT	0.924	0.937	0.750	0.835
HN-ATT	**0.927**	**0.940**	0.750	**0.837**
SNNN	0.918	0.928	0.752	0.827
LBSA	0.922	**0.941**	0.762	0.832
CSNN (200)	0.921	0.938	**0.768**	0.833
CSNN (100)	0.914	0.937	**0.762**	0.835
CSNN (50)	0.916	0.939	**0.765**	0.833

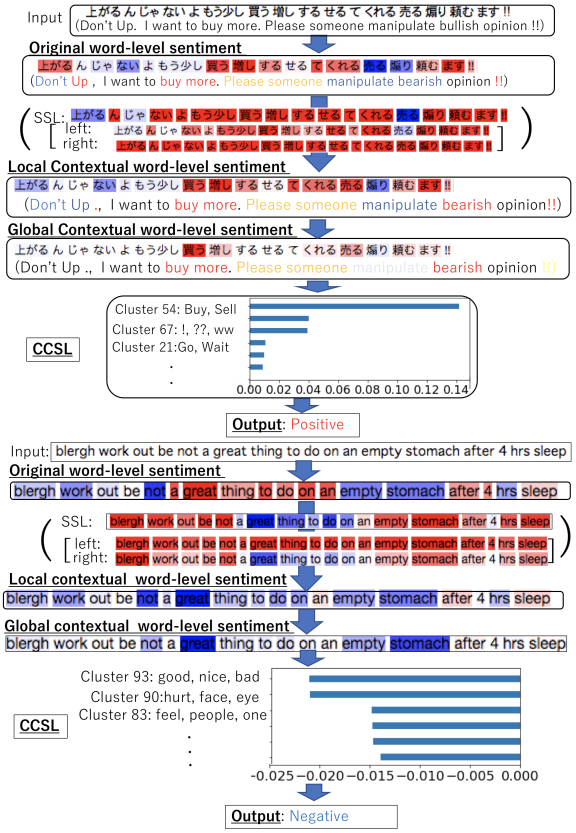

Fig. 3. Local Sentiment Text-visualization Example. Left: Yahoo review and right: Sentiment 140. The color and depth of terms mean polarity (red: > 0 and blue: < 0) and scale of word-level sentiments in each layer.

IP learning. Using several textual datasets, we experimentally demosntrate that 1) IP learning is evvective for improving the explainability of CSNN, and that 2) both the explanation ability and predictability of the CSNN are high. In the future, we will apply this CSNN to documents in other domains or languages. See the supplementary material for the details including theoretical analysis in bit.ly/CSNN20190606.

ACKNOWLEDGMENT

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

REFERENCES

[1] Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP 2014*, 2014.
[2] S. Li, Z. Wang, S. Y. M. Lee, and C.-R. Huang, "Sentiment classification with polarity shifting detection," in *IALP 2013*, 2013, pp. 129–132.
[3] M. Schulder, M. Wiegand, J. Ruppenhofer, and B. Roth, "Towards bootstrapping a polarity shifter lexicon using linguistic features," in *IJCNLP 2017*, 2017, pp. 624–633.
[4] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM-14*, 2014.
[5] Q. Z. X. H. Y. Zou, T. Gui, "A lexicon-based supervised attention model for neural sentiment analysis," in *COLING 2018*, 2018.
[6] Z. Quanshi, Y. N. Wu, and S. C. Zhu, "Interpretable convolutional neural networks," in *CVPR 2018*, 2018.
[7] D. T. Vo and Y. Zhang, "Don't count, predict! an automatic approach to learning sentiment lexicons for short text," in *ACL 2016*, 2016.
[8] Q. Li, "Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits," in *CoNLL 2017*, 2017, pp. 301–310.
[9] T. Ito, H. Sakaji, K. Tsubouchi, K. Izumi, and T. Yamashita, "Text-visualizing neural network model: Understanding online financial textual data," in *PAKDD 2018*, 2018.
[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek, "On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2017.
[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.
[12] S. Karen, V. Andrea, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv:1312.6034*, 2013.
[13] Y. Hechtlinger, "Interpretation of prediction models using the input gradient," in *arXiv:1611.07634*, 2016.
[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, in *Striving for simplicity: The all convolutional net*. ICLR Workshop, 2015.
[15] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS 2013*, 2013.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS 2017*, 2016.
[18] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated self-matching networks for reading comprehension and question answering," in *ACL 2017*, 2017.
[19] M. K. K. Hornik, I. Feinerer and C. Buchta, "Spherical k-means clustering," *Journal of Statistical Software*, vol. 50, no. 10, pp. 1–22, 2012.
[20] P. Nakov, S. Rosenthal, . Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "Semeval-2013 task 2: Sentiment analysis in twitter," in *SemEval 2013*, 2013.
[21] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov, in *SemEval 2014*, 2014.
[22] S. Mohammad, S. Kiritchenko, and X. D. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," in *SemEval-2013*, 2013.
[23] F. Fancellu, A. Lopez, and B. Webber, "Neural networks for negation scope detection," in *ACL 2016*, 2016.
[24] S. Li, S. Yat, M. Lee, Y. Chen, C. R. Huang, and G. Wang, "Sentiment classification and polarity shifting," in *COLING 2010*, 2010.
[25] L. Arras, G. Montavon, K. R. Muller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *EMNLP Workshop*, 2017.
[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *NAACL 2016*, 2016.
[27] Q. Hu, J. Zhou, Q. Chen, and L. He, "Snnn: Promoting word sentiment and negation in neural sentiment classification," in *AAAI 2018*, 2018.
[28] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *KDD*, 2016.
[29] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *ICML*, 2017.