Sentiment Classification with Word Attention based on Weakly Supervised Learning with a Convolutional Neural Network

Gichang Lee [1] **Jaeyun Jeong** [1] **Seungwan Seo** [1] **CzangYeob Kim** [1] **Pilsung Kang** [1]

Abstract

In order to maximize the applicability of sentiment analysis results, it is necessary to not only classify the overall sentiment (positive/negative) of a given document but also to identify the main words that contribute to the classification. However, most datasets for sentiment analysis only have the sentiment label for each document or sentence. In other words, there is no information about which words play an important role in sentiment classification. In this paper, we propose a method for identifying key words discriminating positive and negative sentences by using a weakly supervised learning method based on a convolutional neural network (CNN). In our model, each word is represented as a continuous-valued vector and each sentence is represented as a matrix whose rows correspond to the word vector used in the sentence. Then, the CNN model is trained using these sentence matrices as inputs and the sentiment labels as the output. Once the CNN model is trained, we implement the word attention mechanism that identifies high-contributing words to classification results with a class activation map, using the weights from the fully connected layer at the end of the learned CNN model. In order to verify the proposed methodology, we evaluated the classification accuracy and inclusion rate of polarity words using two movie review datasets. Experimental result show that the proposed model can not only correctly classify the sentence polarity but also successfully identify the corresponding words with high polarity scores.

Keywords: *Weakly Supervised Learning, Word Attention, Convolutional Neural Network, Class Activation Mapping*

1. Introduction

Sentiment analysis and opinion mining is a field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing (NLP) and has also been widely studied in data mining, Web mining, and text mining (Medhat et al., 2014; Liu, 2012; Pang et al., 2008; Ravi & Ravi, 2015) Application domains for sentiment analysis include analyses of customer response to new products or services, analyses of public opinion towards the government's new policies or political issues under debate, etc. (Jo, 2012). In response to increasing needs in diverse domains, various sentiment analysis techniques have been developed (Gui et al., 2017; Cho et al., 2014; Poria et al., 2016; Xianghua et al., 2013; Socher et al., 2013; Kalchbrenner et al., 2014; Tai et al., 2015). However, many of the current sentiment analysis techniques suffer from the over-abstraction problem (Nasukawa & Yi, 2003); the only information obtained from these techniques is the polarity of the document, i.e., whether the nuance of the document is positive or negative. It is difficult to receive more in-depth sentiment analysis results, such as identifying the main words contributing to the polarity classification or finding opposite words or phrase to the overall sentiment of the document, i.e., negative words/phrases in a positive document or positive words/phrases in a negative document.

Recently, attention models have been highlighted in the field of computer vision because of its ability to focus on semantically significant areas in a given image to solve the task of object classification, localization, and detection (Ba et al., 2014; Russakovsky et al., 2015; Mnih et al., 2014). They have also been widely adopted in the field of NLP, as attention models can provide more fruitful interpretations for text analysis tasks (Luong et al., 2015; Shen & Huang,

[1] School of Industrial Management Engineering, Korea University, Seoul, South Korea. Correspondence to: Pilsung Kang <pilsung_kang@korea.ac.kr>.

2016; Rush et al., 2015). Attention models help the NLP model focus on salient words/phrases and transfer these attentions to other machine learning models to solve more complicated tasks such as image captioning or text to image generation (Xu et al., 2015). In addition, as one of the basic building blocks of artificial intelligence (AI) is to understand a human speaker's intention, global technology leaders have released their own AI speakers, such as Amazon's "Eco," Google's "Google Home," and Apple's "Homepod," to collect real-word conversational data in order to upgrade their AI engines. As these AI speakers process the human speaker's query at a sentence level, it becomes more critical to correctly identify the main intentions (words/phrases) of the speaker, which is the ultimate goal of attention models.

It is not that easy to implement an attention model in NLP tasks. This is mainly because most text datasets have document-level labels, i.e., whether the overall nuance of the document is positive or negative, but phrase- or word-level sentiment labels are rarely available. It implies that there is a restriction that the model should learn attention scores for words or phrases without actual labels. To overcome this problem, previous studies modified the structure of a recurrent neural network (RNN) such that the added weights play an attention role inside the model. Applications of RNN-based attention models include document classification (Yang et al., 2016), parsing (Vinyals et al., 2015), machine translation (Bahdanau et al., 2014; Luong et al., 2015), and image captioning (Xu et al., 2015).

In this paper, we propose a sentiment classification with a word attention model based on weakly supervised leaning with a convolutional neural network (CNN), named CAM^2: Classification and Attention Model with a Class Activation Map. The main advantage of the proposed model is its ability to identify crucial words or phrases in a sentence for the sentiment classification perspective without explicit word- or phrase-level sentiment polarity information. It identifies the words by weak labels only, i.e., the sentence-level polarity that is more abstracted but easily available. In the proposed model, words are embedded in a fixed-size of continuous vector space using Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2016). Sentences are represented in a matrix form, whose rows correspond to word vectors, and they are used as the input of a CNN model. The CNN model is trained by considering the sentence-level sentiment polarity as the target, and it produces both the sentence-level polarity score and word-level polarity scores for all words in the sentence, which helps us understand the result of sentence-level sentiment classification. Unlike the existing attention models based on RNN, there is no need to separately learn the weights for the attention. Considering that the same word is used in different contexts for different domains, it is relatively easy to build a dictionary that reflects the characteristics of each domain by using the proposed model.

The rest of this paper is organized as follows. In Section 2, we briefly review and discuss some related works. In Section 3, we demonstrate the architecture of the proposed model. Detailed experimental settings are demonstrated in Section 4 followed by the analysis and discussion of the results. Finally, in Section 5 we present our conclusions.

2. Related Work

In this section, we briefly review the representative studies on for CNN-based document classification (Kim, 2014), weakly supervised learning for CNN-based object detection (Oquab et al., 2015; Zhou et al., 2016), and the RNN-based document attention model named the hierarchical attention network (Yang et al., 2016).

2.1. Convolutional Neural Networks for Document Classification

Kim (2014) showed CNN, which is the most successful neural network structure for image processing, can also work well for text data, especially for document classification. The architecture of Kim (2014) is shown in Figure 1, and it has the following three main ideas:

(1) A large number of filters are used, but the network is not as deep as popular CNN architectures for image processing.

(2) The size of the CNN filter is matched with the vector size of input words.

(3) Multi-channels consisting of static and non-static input vectors are combined.

Experimental results show that the CNN-based document classification model achieved higher classification accuracies than the conventional machine learning-based models, such as the support vector machine or conditional random field, and other deep neural network structures, such as the deep feedforward neural network or recursive neural network. In addition, the word vector could also be customized for a given corpus, and it sometimes yielded better classification performance than pre-trained word vectors.

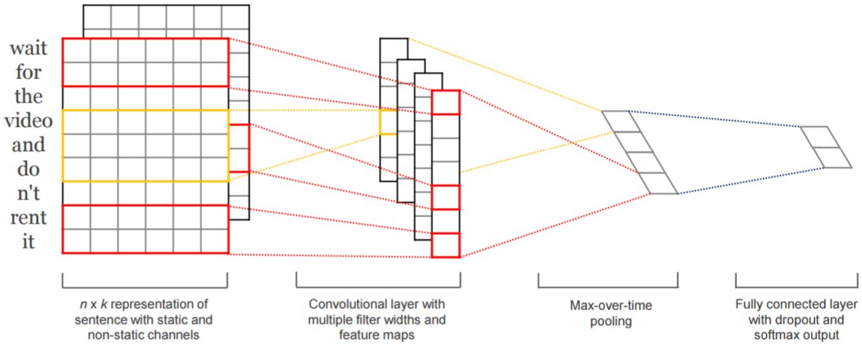
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 1. Model architecture with two channels for an example sentence (Kim, 2014).

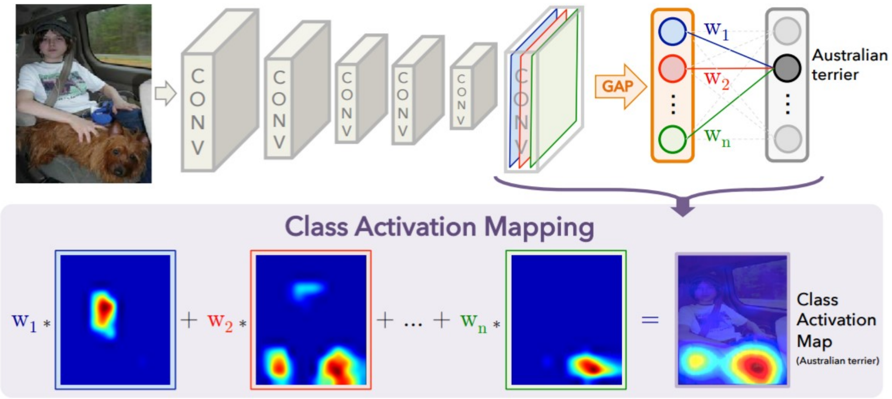

Figure 2. Class activation mapping (Zhou et al., 2016).

2.2. Class Activation Mapping

Oquab et al. (2015) proposed a weakly supervised learning method for object detection without bounding box information. In this study, a standard CNN architecture with max pooling between the final convolution and the output layer was utilized. Zhou et al. (2016) proved the average pooling is more appropriate for the object detection task than the max pooling. The CNN structure and an example of the attention mechanism are shown in Figure 2. In this model, the CNN is trained to correctly classify the object in the input image. In Figure 2, the target of the given image is “Australian terrier,” but no information on the dog’s position in the input image is available during the training. When the training is complete, the weights in the fully connected layers are used to combine the feature map to emphasize the attention area of the original input image. They called this process class activation mapping (CAM), and by utilizing it, not only can the CNN model determine that the “Australian terror” is in the image, but also this classification is mainly inferred by seeing the bottom right part of the image (red area in the final CAM in Figure 2).

2.3. Hierarchical Attention Network

Yang et al. (2016) proposed a hierarchical RNN architecture, inspired by the fact that the document consists of sentences and the sentences are composed of words. In the study, the authors added attention weights to reflect the importance of each sentence and word. As can be seen in Figure 3, the result of their model is the most similar to what we attempted to do in this study. However, the main differences between their work and this work is that Yang et al. (2016) employed an RNN as the base model and the attention weights were separately learned from the corpus. However, a CNN is employed as the base model for sentiment classification in this study, and we do not explicitly train the model to learn the word-level attention scores.

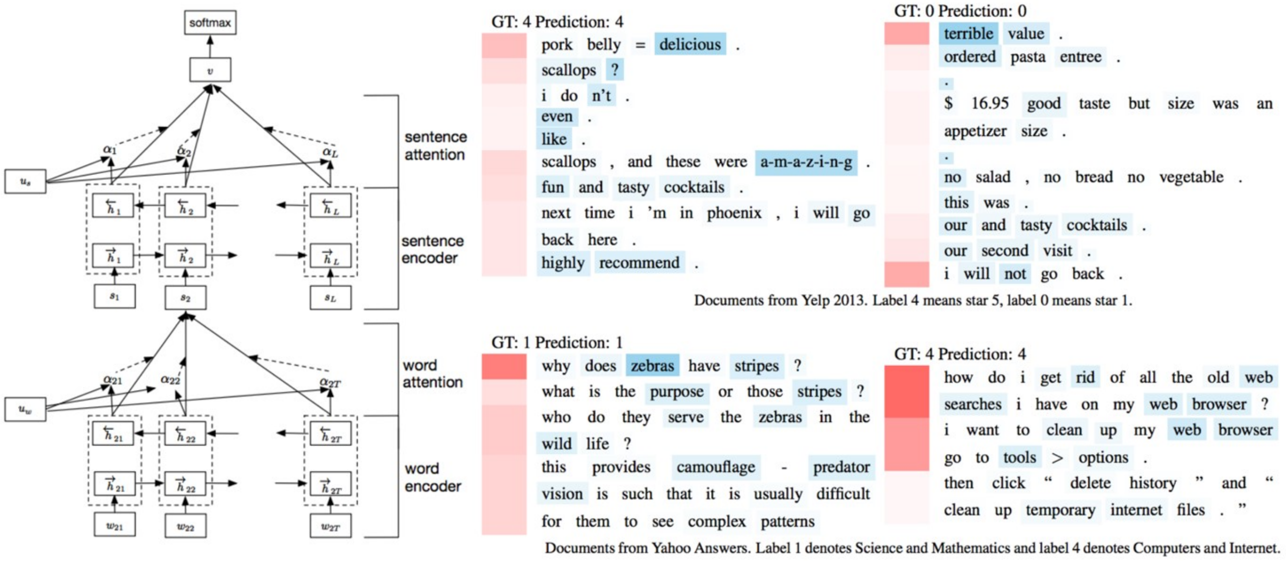
softmax
v
sentence attention
sentence encoder
word attention
word encoder
u_s
u_w
α_1
α_2
α_L
α_{21}
α_{22}
α_{2T}
s_1
s_2
s_L
w_{21}
w_{22}
w_{2T}
GT: 4 Prediction: 4
pork belly = delicious .
scallops ?
i do n't .
even .
like .
scallops , and these were a-m-a-z-i-n-g .
fun and tasty cocktails .
next time i 'm in phoenix , i will go
back here .
highly recommend .
GT: 0 Prediction: 0
terrible value .
ordered pasta entree .
.
$ 16.95 good taste but size was an
appetizer size .
.
no salad , no bread no vegetable .
this was .
our and tasty cocktails .
our second visit .
i will not go back .
Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.
GT: 1 Prediction: 1
why does zebras have stripes ?
what is the purpose or those stripes ?
who do they serve the zebras in the
wild life ?
this provides camouflage - predator
vision is such that it is usually difficult
for them to see complex patterns
GT: 4 Prediction: 4
how do i get rid of all the old web
searches i have on my web browser ?
i want to clean up my web browser
go to tools > options .
then click " delete history " and "
clean up temporary internet files . "
Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.

Figure 3. Hierarchical Attention Network (Yang et al., 2016).

3. Classification and Attention Model based on Class Activation Map: CAM^2

3.1. Overall Framework

Figure 4 shows the overall framework of the proposed method. After collecting the sentences, low-level embedding is performed by the Word2Vec, GloVe, and FastText methods, and the word vectors in the sentence are concatenated to form the initial input matrix for the CNN. Once the CNN model training is completed, the polarity of a given test sentence is predicted. Then, the weights of the fully connected layer are used to combine the feature maps to produce the attention score for every single word in the sentence.

3.2. Network Architecture

The architecture of the CNN used in this paper is basically rooted in the CNN architecture used in Kim (2014). However, since the CNN used in Kim (2014) was originally designed for document classification, we made some modifications to it to facilitate the extraction of essential words or phrases. First, the zero-padding is added before the first word and after the last word in the sentence to make that the number of times that each word is included in the receptive field during convolution the same, irrespective of the word's position in the sentence. Second, we applied average-pooling instead of max-pooling. According to Zhou et al. (2016), average-pooling and max-pooling are essentially similar, but using average-pooling is advantageous in identifying the overall scope of the target. Third, we increased the number of filters compared to the CAMs used in Oquab et al. (2015) and Zhou et al. (2016). As these CAMs are specialized for image processing, the receptive field of convolution is a square (ex: 3 × 3). However, the receptive field of the proposed CAM^2 is a rectangular (ex: 3 × word embedding dimension), which integrates a larger amount of information in one scalar value compared to the convolutional filter in image processing. To prevent a possible loss of information due to a larger receptive field, we used a much larger number of convolution filters than was used in (Kim, 2014). Finally, we used more various word embedding techniques to form an input matrix of a sentence. Kim (2014) only used the Word2Vec for word embedding, but we consider two recently developed word embedding techniques: GloVe and FastText.

3.3. Classification and Attention Model based on Class Activation Map

The input of CNN, $\mathbf{x}_{1:l}$ is created by concatenating the word vectors in a sentence and zero-paddings. We used four type of inputs *CNN-rand, CNN-static, CNN-non-static,* and *CNN-Multichannel*. The *CNN-rand* uses a randomly initialized word vector while *CNN-static* and *CNN-non-static* use the word vectors pre-trained by the Word2Vec. *CNN-Multichannel* uses the word vectors pre-trained by the Word2Vec, GloVe, and FastText. Let k, d, and h denote the dimension of the word embedding vector, number of maximum words in a sentence, and the height of the receptive

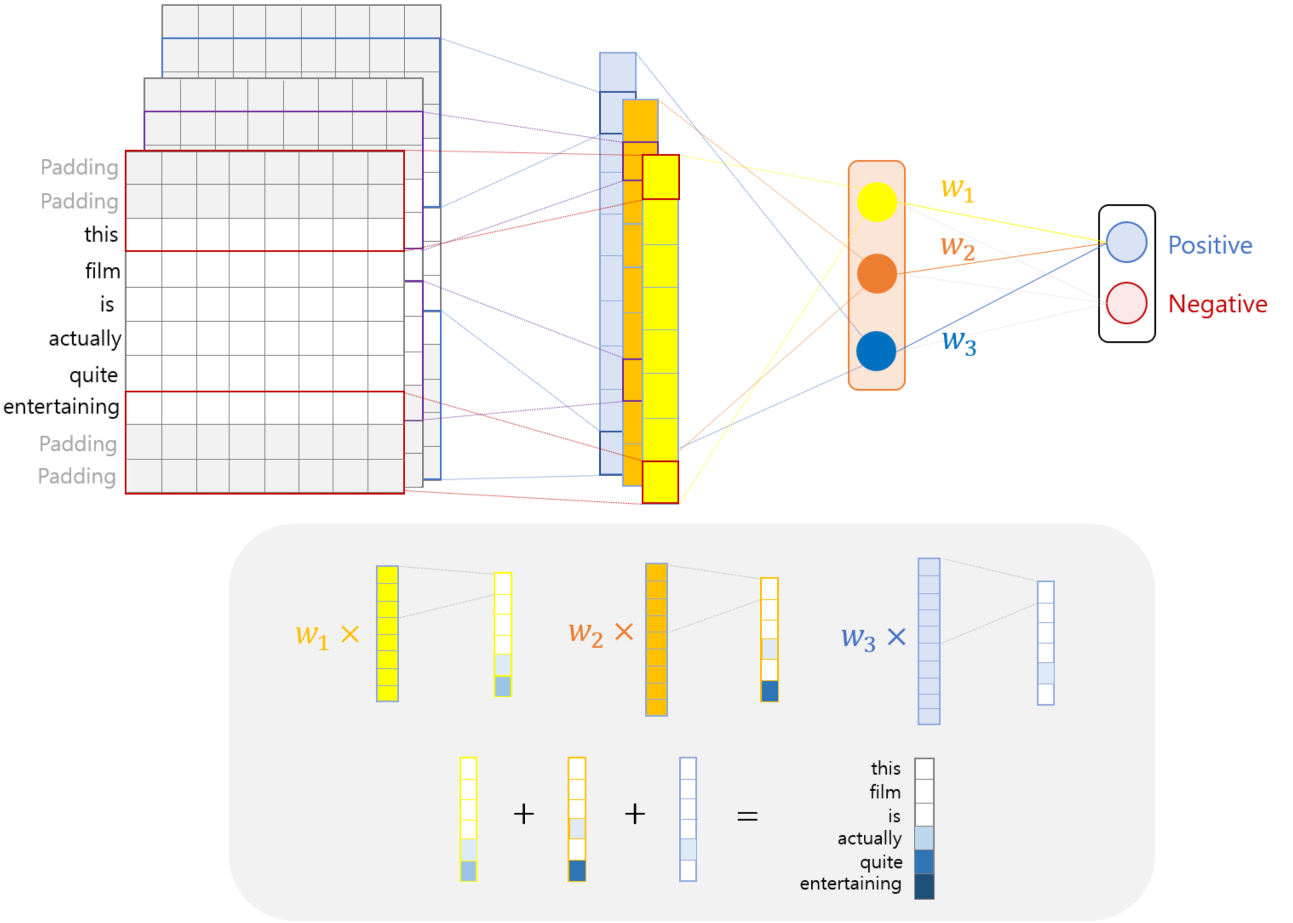

Figure 4. Framework of proposed method.

field of convolution, respectively, then the input matrix $\mathbf{X} \in R^{([d+2(h-1)]k)}$ is constructed as follows. The zero-padding is first performed before and after $\mathbf{x}_{1:d}$ so that the number of times that each word is included in the receptive field during convolution is the same (h times).

$$\mathbf{X} = \mathbf{x}_{1:l} = \underbrace{0 \oplus \ldots \oplus 0}_{h-1} \oplus \underbrace{\mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_d}_{d} \oplus \underbrace{0 \oplus \ldots \oplus 0}_{h-1}. \tag{1}$$

When the window size of the CNN filter, i.e., the height of filter is h, the i-th feature map f_i is constructed as follows. As the size of CNN filter $\mathbf{w}$ is $h \times d$ and zero-padding is performed in the previous step, f_i becomes a I-dimensional vector, where I is $(d + h - 1)$.

$$\mathbf{f_i} = [f_{1i},\ f_{2i},\ \ldots,\ f_{li}]^T, \tag{2}$$
$$f_{ji} = ReLu(\mathbf{W}_{conv} \odot \mathbf{x}_{j:j+h-1} + b), \tag{3}$$
$$\mathbf{W}_{conv} \in R^{h \times k},\ b \in R. \tag{4}$$

Let $\hat{f}_l$ be the scalar value computed by applying the average pooling to the feature map $\mathbf{f}_i$. The final feature vector $\mathbf{z}$ passed to the fully connected layer is constructed as follows. Considering that n feature maps are computed for a given sentence, $\mathbf{z}$ becomes an n-dimensional vector.

$$\mathbf{z} = [\hat{f}_1,\ \hat{f}_2,\ \ldots,\ \hat{f}_n]^T, \tag{5}$$

where n is n_{ftypes} (the number of filter type) $\times$ $n_{filters}$ (the number of filters for each type). The output of the fully connected layer for the i-th sentence is $\mathbf{y}$,

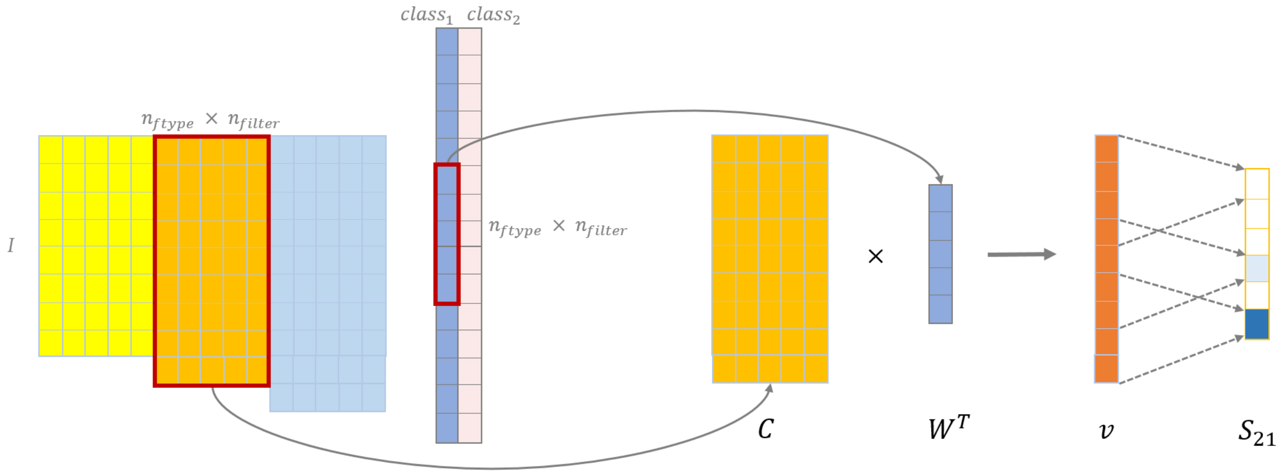

Figure 5. An example of computing a score vector.

computed as follows:

$$\mathbf{y} = \mathbf{W}_{fc} \cdot \mathbf{z} + \mathbf{b}_{fc}, \tag{6}$$

$$\mathbf{W}_{fc} \in R^{c \times n}, \tag{7}$$

$$\mathbf{b}_{fc} \in \; R^{c}, \tag{8}$$

$$c: \text{ the number of classes.}$$

Once the CNN model is trained, the sentiment importance score of each word is computed as follows. An illustrated example of the following process is provided in Figure 5. Let F_l be the feature maps corresponding to the l-th filter type and w_{lc_i} be the row vector of $\mathbf{W}_{fc}$ for the l-th filter type and the c_i-th class. Then, the score vector $\mathbf{v}$ is computed as

$$\mathbf{v} = \mathbf{F}_l \cdot \mathbf{w}_{lc_i}^T, \tag{9}$$

$$\mathbf{F}_l \in R^{I \times n_{filter}}, \tag{10}$$

$$\mathbf{w}_{lc_i}^T \in R^{nfilter}. \tag{11}$$

The p-th element in the score vector s_{lc_i} corresponding to the l-th filter type and the c_i-th class is computed by averaging h elements with the step size of 1, which makes the $\mathbf{s}_{lc_i}$ a d-dimensional vector, regardless of the height of filters:

$$\mathbf{s}_{lc_i} = \frac{1}{h} \sum_{q=p}^{p+h-1} \mathbf{V}_q. \tag{12}$$

The final sentiment score of the words in the sentence to c_i-th class, $\mathrm{CAM}_{c_i}^2$ is computed by

$$\mathrm{CAM}_{c_i}^2 = \sum_{l=1}^{n_{ftype}} \mathbf{s}_{lc_i}. \tag{13}$$

3.4. Word Embedding

We employed four different word embedding methods to construct the input matrix X: random vectors, Word2Vec, GloVe, and FastText. With the random vectors, the elements of the word vectors were randomly initialized, and they were updated during the CNN training. For the latter three methods, word embedding vectors were separately trained using the same corpus for sentiment classification. We also compared the static word embedding and non-static word embedding methods for CAM^2 according to whether the word embedding vectors are updated during the CNN training (non-static) or not (static). In addition, two multi-channel input matrices were also considered. In summary, we tested the following five input matrices for CAM^2.

(1) *CNN-Rand*: word vectors are randomly initialized and they are updated during the CNN training.

(2) *CNN-Static*: word vectors are trained by Word2Vec. They are not updated during the CNN training.

(3) *CNN-Non-Static*: word vectors are trained by Word2Vec first, and they are updated during the CNN training.

(4) *CNN-2ch*: *CNN-Static* and *CNN-Non-Static* are combined. The input of CNN becomes a 3-dimensional (I × k × 2) tensor.

(5) *CNN-4ch*: Three matrices with word vectors trained by Word2Vec, GloVe, and FastText are used. They are updated during the CNN training. The *CNN-Non-Static* method is used as the

Table 1. Rating distributions of the IMDB dataset

Score	1	2	3	4	7	8	9	10
Reviews	10,122	4,586	4,961	5,531	4,803	5,859	4,607	9,731
Class	Negative				Positive			

Table 2. Rating distributions of the WATCHA dataset

score	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Reviews	50,660	66,184	62,094	163,272	173,650	411,757	424,378	652,250	297,327	416,096
Class	Negative				Not used					Positive

Table 3. The number of tokens

IMDB	WATCHA
115,205	424,027

Table 4. The hyper-parameters of the CNN

Filter type (window size)	3 (tri-gram) 4 (quad-gram) 5 (5-gram)
N. filters	128 each
Doc. length	100 words
Dropout rate	0.5
L_2 regularization (λ)	0.1
Batch size	64

fourth matrix. The input of CNN becomes a 3-dimensional (I × k × 4) tensor.

Table 5. The test accuracy between methodology

Test	IMDB	WATCHA
CNN-Rand	0.8435	0.7793
CNN-Static	0.7750	0.7150
CNN-Non-Static	0.8257	0.7538
CNN-2channel	0.8300	0.7602
CNN-4channel	0.8729	0.7533

Table 6. CAM example

Word	Score
this	0.0145
film	0.0291
is	0.1324
actually	0.2183
quite	0.2561
entertaining	**0.3496**

4. Experimental Settings

4.1. Data Sets & Target Labeling

To verify the proposed CAM^2, we used two sets of movie reviews, one written in English and the other written in Korean. Not only do movie reviews have explicit sentiment labels (ratings or stars), but they generally also have more subjective expressions compared to other formal texts such as news articles. For the English movie review dataset, we used the publicly available IMDB dataset (Maas et al., 2011), while Korean movie reviews were collected directly from the WATCHA website (https://watcha.net), which is the largest movie recommendation service in Korea. Each dataset consists of review sentences and ratings. The distributions of ratings for the IMDB and WATCHA are shown in Table 1 and 2.

As shown in Table 2, the ratings are well-balanced in the IMDB dataset. Hence, we used the reviews with ratings smaller than or equal to 4 as negative examples, whereas the reviews with ratings greater than or equal to 7 were used as positive examples. Unlike for the IMBD dataset, the ratings of the WATCH dataset are highly skewed toward the positive scores. Therefore, we used the reviews with ratings smaller than or equal to 2 as negative examples whereas only the reviews with 5-point-ratings were used as positive examples. In both datasets, 70% of the reviews were used as training data, and the remaining 30% were used as test data.

4.2. Word Embedding, CNN Parameters, and Performance Measure

Each sentence was split into tokens using the space. The punctuations and numbers were removed. All tokens were used to learn the word embedding vectors. We fixed the dimension of word embedding to 100 and set the window size of Word2Vec and FastText to 3.

Table 7. Frequently appeared words in the positive/negative sentences in in the IMDB test dataset (semantically positive or negative words are colored in blue and red, respectively)

Positive					**Negative**				
CNN-Rand	*CNN-Static*	*CNN-Non-Static*	*CNN-2channel*	*CNN-4channel*	*CNN-Rand*	*CNN-Static*	*CNN-Non-Static*	*CNN-2channel*	*CNN-4channel*
the	and	and	and	and	the	the	the	the	the
and	great	is	is	is	a	is	and	and	and
a	is	the	the	a	and	was	worst	worst	of
of	a	a	a	the	of	and	of	of	a
is	very	of	of	of	to	bad	a	a	worst
to	the	s	s	s	is	a	is	is	is
I	well	excellent	excellent	it	I	this	the	was	was
in	film	it	great	excellent	in	of	awful	awful	I
this	of	great	it	to	this	plot	boring	to	to
it	it	to	in	I	that	just	was	I	awful
that	I	in	to	great	was	acting	to	boring	this
was	as	an	I	it	it	movie	I	movie	movie
as	excellent	I	an	in	movie	I	bad	this	boring
movie	wonderful	was	perfect	perfect	for	awful	poor	poor	bad
with	movie	perfect	with	very	with	script	this	bad	s
for	story	as	as	was	as	boring	movie	waste	in
film	in	best	very	fun	have	to	waste	terrible	waste
but	favorite	very	best	by	on	that	terrible	in	poor
on	beautiful	enjoyed	enjoyed	enjoyed	film	so	in	s	terrible
an	my	with	wonderful	an	but	t	s	with	for
have	good	wonderful	fun	as	not	terrible	horrible	are	with
are	comedy	fun	by	with	be	it	are	as	as
one	loved	by	movie	best	are	stupid	with	by	are
his	also	amazing	amazing	wonderful	you	horrible	for	acting	it
you	most	movie	loved	loved	an	in	by	for	that
not	s	most	that	amazing	at	are	film	horrible	film
be	best	loved	superb	most	his	film	as	film	by
who	enjoyed	superb	film	are	one	no	it	that	horrible
by	love	that	most	for	from	worst	poorly	it	so

For Word2Vec and FastText, we used the skip-gram structure, while unigram was used to create the co-occurrence matrix for GloVe. The total number of tokens for each dataset is shown in Table 3.

The hyper-parameters for training CNN are summarized in Table 4. We used three different window sizes (how many words are considered in one receptive field), while the number of filters was fixed to 128. The document length, i.e., the maximum number of words, was set to 100. For sentences shorter than 100 words, zero-paddings were added after the last word, whereas the last words were trimmed if sentences were longer than 100 words. We also used two regularization methods. The dropout is an implicit regularization that ignores some weights in each step (dropout rate = 0.5 in this study), whereas the L_2 regularization is an explicit regularization that adds the L_2-norm of the total weight in the loss function.

5. Result

5.1. Classification Performance

Table 5 shows the classification accuracies for the five CNN models. It is worth noting that the CNN-Static

Table 8. Frequently appearing words in the positive/negative sentences in in the WATCHA test dataset (semantically positive or negative words are in blue and red fonts, respectively)

Positive					**Negative**				
CNN-Rand	*CNN-Static*	*CNN-Non-Static*	*CNN-2channel*	*CNN-4channel*	*CNN-Rand*	*CNN-Static*	*CNN-Non-Static*	*CNN-2channel*	*CNN-4channel*
영화	영화	영화	영화	영화	영화	영화	영화	영화	영화
너무	수	최고의 (best)	최고의 (best)	너무	너무	너무	너무	너무	너무
이	최고의 (best)	너무	너무	최고의 (best)	이	왜	이	그냥	그냥
수	정말	수	그리고	수	더	없고 (none)	그냥	왜	왜
왜	다시	그리고	그	그	왜	없는 (none)	왜	이	없는 (none)
영화를	잘	정말	수	그리고	수	그냥	더	없는 (none)	이
없는 (none)	그	그	가장	정말	없다 (none)	이	그	그	없고 (none)
그냥	그리고	가장	정말	더	그냥	좀	없는 (none)	없고 (none)	다
없다 (none)	또	더	더	가장	영화를	영화는	영화는	더	그나마
더	진짜	최고 (best)	이	이	그	없다 (none)	영화를	다	그
다	너무	있는	최고 (best)	잘	없는 (none)	이런	수	정말	더
이런	가장	진짜	진짜	다시	다	느낌	없고 (none)	없다 (none)	정말
것	내	잘	다시	최고 (best)	것	뻔한 (obvious)	좀	영화를	영화를
그	있는	것	잘	진짜	내가	영화를	다	그나마	좀
영화가	좋다 (good)	이	것	것	이런	보는	한	좀	영화는
영화는	아름다운 (beautiful)	모든	영화가	보고	영화가	안 (not)	정말	영화는	없다 (none)
진짜	함께	영화가	보고	좋다 (good)	정말	내	없다 (none)	내	이런
본	더	보고	있는	봐도	영화는	그	그나마	수	수
좀	이	내	내가	내	이렇게	무슨	보는	이런	별
정말	최고 (best)	다시	마지막	있는	좀	건	이런	대한	별로 (not much of)
내	작품	내가	내	본	보는	것도	내	보는	영화가
잘	모든	봐도	모든	모든	본	스토리	건	차라리 (rather)	내
내가	내가	본	한	완벽한 (perfect)	진짜	많이	대한	한	차라리 (rather)
이렇게	본	내내	좋다 (good)	이렇게	잘	차라리 (rather)	차라리 (rather)	별	안
한	봐도	또	완벽한 (perfect)	내가	없고 (none)	것	잘	건	아깝다 (wasted)
보는	중	좋다 (good)	봐도	하는	한	아닌 (not)	별	잘	느낌
이건	있을까	마지막	영화를	다	내	듯	안	뻔한 (obvious)	대한
좋은 (good)	꼭	대한	다	또	스토리	뭘	내가	봤는데	최악의 (worst)
보고	모두	완벽한 (perfect)	본	한	이건	못한 (not)	영화가	최악의 (worst)	것

Table 9. Example of word attention for a positively classified sentence in the IMDB dataset

Methodology	Sentence
Raw text	I'm normally not a Drama/Feel good movie kind of guy but once I saw the trailer for Radio I couldn't resist. Not only is this a great film but it also has great acting. Cuba Gooding Jr. did an excellent job portraying James Robert Kennedy a.k.a. RAdio. Ed Harris also did a fantastic job as Coach Jones. I was pleasantly surprised to see some comedy in it as well. So for a great story great acting and a little comedy I give Radio a 10 out of 10! (10 / 10 points)
CNN-Rand	I m normally not a Drama Feel good movie kind of guy but once I saw the trailer for Radio I couldn t resist Not only is this a great film but it also has great acting Cuba Gooding Jr did an excellent **job portraying James Robert** Kennedy a k a RAdio Ed Harris also did a fantastic job as Coach Jones **I was pleasantly** surprised to see some **comedy in** it as well So for a great story great acting and a little comedy I give Radio a out of Positive
CNN-Static	I m normally not a Drama Feel good movie kind of guy but once I saw the trailer for Radio I couldn t resist Not only is this a **great film** but it also has great acting Cuba Gooding Jr did an **excellent** job portraying James Robert Kennedy a k a RAdio Ed Harris also did a **fantastic** job as Coach Jones I was pleasantly surprised to see some comedy in it as well So for **a great story great acting** and a little comedy I give Radio a out of Positive
CNN-Non-Static	I m normally not a Drama Feel good movie kind of guy but once I saw the trailer for Radio I couldn t resist Not only is this a great film but it also has great acting Cuba Gooding Jr did **an excellent job** portraying James Robert Kennedy a k a RAdio Ed Harris also did a **fantastic job as** Coach Jones I **was pleasantly surprised** to see some comedy in it as well So for a great story great acting and a little comedy I give Radio a out of Positive
CNN-2channel	I m normally not a Drama Feel good movie kind of guy but once I saw the trailer for Radio I couldn t resist Not only is this a great film but it also has great acting Cuba Gooding Jr did **an excellent job** portraying James Robert Kennedy a k a RAdio Ed Harris also did a **fantastic job as** Coach Jones I **was pleasantly surprised** to see some comedy in it as well So for a great story great acting and a little comedy I give Radio a out of Positive
CNN-4channel	I m normally not a Drama Feel good movie kind of guy but once I saw the trailer **for Radio** I couldn t resist Not only is this a great film but it also has great acting Cuba Gooding Jr did an **excellent job** portraying James Robert Kennedy a k a RAdio Ed Harris also did a **fantastic job** as Coach Jones I **was pleasantly surprised** to see some comedy in it as well So for a great story great acting and a little comedy I give Radio a out of Positive

resulted in the lowest classification accuracy for both IMDB and WATCHA datasets. Since the CNN-Static is the only model which does not update the word embedding vectors during the CNN training, updating the word embedding vectors for a given corpus during the model training, whether or not the word vectors are independently trained before, is encouraged to achieve better classification performance.

Table 6 shows an example of CAM^2 for a test sentence. The overall sentiment of this sentence is classified as positive. For each word, the higher the score, the CNN model considers it as a significantly contributing word to the overall sentiment. Thus, the word 'entertaining' had the greatest impact on the classification of this review as being positive.

5.2. Finding Sentimental Words

Table 7 provides the frequent words listed in the IMDB test dataset by selecting the top five highly scored words in the sentences classified as positive (left five columns) and negative (right five columns). It is worth noting that although the CNN-Rand yielded a relatively good classification performance compared to other techniques, it identified the least emotional words among the five CNN models. Although the classification performance of CNN-Static was the worst, its attention mechanism seemed to work well, in that many emotional words were highly ranked. In terms of classification performance, it is important whether or not the input vector is updated in the training process. However, for the sake of word attention in sentiment

Table 10. Example of word attention for a negatively classified sentence in the IMDB dataset

Methodology	Sentence
Raw text	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If he was the Rain Man it would've been fine but he's not. (3/10 points)
CNN-Rand	This is one of the most boring films I've ever seen The three main cast members just didn t seem to click well Giovanni Ribisi s character was quite annoying For some reason he seems to like repeating what he says If he was the Rain Man it **would ve been fine but he s** not **Negative**
CNN-Static	This is one of the most **boring** films I ve ever seen The three main cast members just didn t seem to click well Giovanni Ribisi s character was **quite annoying For some reason** he seems to like repeating what he says If he was the Rain Man it would ve been fine but he's not **Negative**
CNN-Non-Static	This is one of the **most boring films** I've ever seen The three main cast members just didn t seem to click well Giovanni Ribisi s character was quite **annoying For some** reason he seems to like repeating what he says If he was the Rain Man it would ve been fine but he's not **Negative**
CNN-2channel	This is one of the **most boring films** I ve ever seen The three main cast members just didn t seem to click well Giovanni Ribisi s character was quite **annoying For** some reason he seems to like repeating what he says If he was the Rain Man it would ve been fine but he s not **Negative**
CNN-4channel	This is one of the **most boring** films I ve ever seen The three main cast members just didn t seem to click well Giovanni Ribisi s **character was quite annoying** For some reason he seems to like repeating what he says If he was the Rain Man it would ve been fine but he s not **Negative**

classification, it becomes more important whether the general grammatical relationship between the words are well-preserved in the word embedding vector (not updated for classification task).

Table 8 provides the frequent words listed in the WATCHA test dataset by selecting the top five highly scored words in the sentences classified as positive (left five columns) and negative (right five columns). In this case, the emotional word in the upper word list is somewhat overlapped with other methods compared to the IMDB dataset. This is because Korean is an agglutinative language, which tends to have a high rate of affixes per word. For example, "없다, 없는, 없고...(none)," "안, 아닌, 못...(not)," and "차라리(rather)" are usually used in Korean for negative expressions. Experimental results confirm that these words are more frequently used in the negative reviews than in the positive reviews (except CNN-Rand).

5.3. Word Attention: IMDB

Table 9 shows an example of word attention of a positively classified sentence in the IMDB dataset. The words highlighted in blue are the top 10% highly scored words in the sentence. The four models except the *CNN-Rand* can successfully capture semantically positive words or phrases (ex. excellent, fantastic, and was pleasantly surprised). In particular, the *CNN-Static* is especially good at paying attention to longer sentimental phrases such as "a great story great acting."

Table 10 shows an example of word attention of a negatively classified sentence in the IMDB dataset. The words highlighted in red are the top 10% highly scored words in the sentence. If one reads the review, he/she can easily recognize multiple negative expressions within the review, which results in different attention words or phrases according to different models. For example, the *CNN-Non-Static, CNN-2channel,* and *CNN-4channel* pay attention to "boring" and "annoying," both of which are clearly negative expressions when used in a movie review. However, there is another explicit negative expression, namely, "it would (have) been fine," which receives an attention by the *CNN-Rand.*

Table 11 shows an example of attention results for a sentence whose predicted class is different according to the CNN models because of mixed emotional expressions within the sentence. In this case, the words in the top 10% highest scores are highlighted in blue and those in the bottom 10% lowest scores are highlighted in red if the sentence is classified as positive. The highlighting scheme is reversed if the sentence is classified as negative. Likewise, the *CNN-Static, CNN-Non-Static, CNN-2channel,* and *CNN-4channel* have relatively better attention performances than the CNN-Rand. Again, the *CNN-Static* has a relatively good performance in capturing longer emotional phrases such as "is also very interesting and touching."

Table 11. Example of word attention for a sentence in the IMDB dataset whose predicted class is different according to CNN models

Methodology	**Sentence**
Raw text	This movie has a lot to recommend it. The paintings the music and David Hewlett's naked butt are all gorgeous! The plot a story of redemption forgiveness and courage in the face of adversity is also very interesting and touching – and it's not predictable which is saying quite a lot about a movie in this day and age. But the acting is mediocre the direction is confusing and the script is just odd. It often felt like it was trying to be a parody but I never figured out what it was trying to be parody *of*. (9 / 10 points)
CNN-Rand	**This movie has a lot** to recommend **it** The paintings the music **and David Hewlett s** naked butt are all gorgeous The plot a story of redemption **forgiveness and courage in** the face of adversity is also very interesting and touching and it s not predictable which is saying quite a lot about a movie in this day and age But **the** acting is mediocre the direction is confusing and the **script is just odd It** often felt like it was trying to be a parody but I never figured out what it was trying to be parody of Negative
CNN-Static	This movie has a lot to recommend it **The paintings the music** and David Hewlett s naked butt are all gorgeous The plot a story of redemption forgiveness and courage in the face of adversity **is also very interesting and touching** and it s not predictable which is saying quite a lot about a movie in this day and age But the acting **is mediocre the direction is confusing and the script is** just odd It often felt like it was trying to be a parody but I never figured out what it was trying to be parody of Negative
CNN-Non-Static	This movie has a lot to recommend it **The paintings** the **music and David Hewlett s** naked butt are all gorgeous The plot a story of redemption forgiveness and courage in the **face** of adversity is also very interesting **and touching** and it s not **predictable which is** saying quite a lot about a movie in this day and age But the **acting is mediocre the direction is** confusing and the script is just odd It often felt like it was trying to be a parody but I never figured out what it was trying to be parody of Positive
CNN-2channel	This movie has a lot to **recommend it The paintings** the music **and David Hewlett** s naked butt are all gorgeous The plot a story of redemption forgiveness and courage in the face of adversity is also very interesting **and touching and** it s not **predictable which is** saying quite a lot about a movie in this day and age But the **acting is mediocre the direction is** **confusing** and the script is just odd It often felt like it was trying to be a parody but I never figured out what it was trying to be parody of Positive
CNN-4channel	This movie has a lot to **recommend it The paintings the music and David Hewlett** s naked butt are all gorgeous The plot a story of redemption forgiveness and courage in the face of adversity is also very interesting and **touching** and it s not **predictable which** is saying quite a lot about a movie in this day and age But the acting **is mediocre the direction is confusing** and the **script** is just odd It often felt like it was trying to be a parody but I never figured out what it was trying to be parody of Positive

5.4. Word Attention: WATCHA

Table 12 shows an example of word attention of a positively classified sentence in the WATCHA dataset. The words highlighted in blue are the top 10% highly scored words in the sentence. In this sentence, there are two obvious positive expressions, i.e., 감탄스럽다 (impressing) and 존경스럽다 (admirable); the former was successfully detected by *CNN-Static, CNN-Non-Static, CNN-2channel,* and *CNN-4channel* while the latter was detected by *CNN-Rand*.

Table 13 shows an example of word attention of a negatively classified sentence in the WATCHA dataset. The words highlighted in blue are the top 10% highly scored words in the sentence. This sentence also has two semantically explicit negative expressions: "불필요하고 의미없는 가오 (unnecessary and meaningless flaunt)" and "한마디로 총체적 난국 (a total crisis in a word)." The *CNN-Rand* focused on the former expression, whereas the rest of the four models focused on the latter expression. Similar to the example of the positive sentence in Table 12, it seems that the atten-

Table 12. Example of word attention for a positively classified sentence in the WATCHA dataset

Methodology	**Sentence**
Raw text	살라딘의 기사도 정신이 진짜 감탄스럽다. 예수상을 다시 세우고 십자가 바닥을 안 밟고 지나가는 장면이 존경스럽다. (5 / 5 points) (Saladin's Chivalry spirit is truly amazing. I'm very impressed by the scene of setting up the Jesus prize and passing without stepping on the floor of the cross.)
CNN-Rand	살라딘의 기사도 정신이 진짜 감탄스럽다 예수상을 다시 세우고 십자가 바닥을 안 **밟고** 지나가는 장면이 **존경스럽다** Positive
CNN-Static	살라딘의 기사도 정신이 진짜 **감탄스럽다 예수상을** 다시 세우고 십자가 바닥을 안 밟고 지나가는 장면이 존경스럽다 Positive
CNN-Non-Static	살라딘의 기사도 정신이 진짜 **감탄스럽다 예수상을** 다시 세우고 십자가 바닥을 안 밟고 지나가는 장면이 존경스럽다 Positive
CNN-2channel	살라딘의 기사도 정신이 진짜 **감탄스럽다 예수상을** 다시 세우고 십자가 바닥을 안 밟고 지나가는 장면이 존경스럽다 Positive
CNN-4channel	살라딘의 기사도 정신이 진짜 **감탄스럽다 예수상을** 다시 세우고 십자가 바닥을 안 밟고 지나가는 장면이 존경스럽다 Positive

Table 13. Example of word attention for a negatively classified sentence in the WATCHA dataset

Methodology	**Sentence**
Raw text	영화 전체를 통틀어 가장 불필요하고 의미없는 가오를 잡는 여자가 환호를 받고 있는 아이러니한 영화! 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 한마디로 총체적 난국. (2 / 5 points) (An ironic movie in which the most unnecessary and meaningless flaunt woman in the whole movie is being cheered! Soundtracks are acceptable but storytelling makes the audience run down. A total impasse in a word.)
CNN-Rand	영화 전체를 통틀어 가장 **불필요하고 의미없는 가오를** 잡는 여자가 환호를 받고 있는 아이러니한 영화 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 한마디로 총체적 난국 **Negative**
CNN-Static	영화 전체를 통틀어 가장 불필요하고 의미없는 가오를 잡는 여자가 환호를 받고 있는 아이러니한 영화 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 **한마디로 총체적 난국** **Negative**
CNN-Non-Static	영화 전체를 통틀어 가장 불필요하고 의미없는 가오를 잡는 여자가 환호를 받고 있는 아이러니한 영화 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 **한마디로 총체적 난국** **Negative**
CNN-2channel	영화 전체를 통틀어 가장 불필요하고 의미없는 가오를 잡는 여자가 환호를 받고 있는 아이러니한 영화 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 **한마디로 총체적 난국** **Negative**
CNN-4channel	영화 전체를 통틀어 가장 불필요하고 의미없는 가오를 잡는 여자가 환호를 받고 있는 아이러니한 영화 사운드트랙은 인정하더라도 관객을 지나가는 메트로폴리스 행인만도 못하게 다루는 스토리텔링 **한마디로 총체적 난국** **Negative**

tion mechanism of *CNN-Rand* is somewhat different from those of the other models. This is mainly because the word embedding vectors are not updated to reflect the user's rating information. Hence, more general

Table 14. Example of word attention for a sentence in the IMDB dataset whose predicted class is different according to CNN models

Methodology	Sentence
Raw text	이렇게 재미없고 그래픽도 꾸지고 난장판인 엑스맨을 과거의 이야기로 새로 시작한 메튜 본 감독과 깔끔하게 다시 재정리한 브라이언 싱어 감독에게 박수를... (1 / 5 points) (I would like to pay tribute to Bryan Singer, who just reconstituted this boring and messy X-Men as a story of the past, and Matthew Vaughn, who neatly rearranged it again.)
CNN-Rand	이렇게 재미없고 그래픽도 꾸지고 난장판인 엑스맨을 과거의 **이야기로** 새로 **시작한** 메튜 본 감독과 깔끔하게 다시 재정리한 브라이언 **싱어** 감독에게 **박수를** **Negative**
CNN-Static	이렇게 **재미없고 그래픽도** 꾸지고 난장판인 엑스맨을 과거의 이야기로 새로 시작한 메튜 본 감독과 깔끔하게 다시 재정리한 브라이언 싱어 **감독에게** **박수를** Positive
CNN-Non-Static	이렇게 **재미없고 그래픽도** 꾸지고 난장판인 엑스맨을 과거의 이야기로 새로 시작한 메튜 본 감독과 **깔끔하게 다시** 재정리한 브라이언 싱어 감독에게 박수를 **Negative**
CNN-2channel	이렇게 **재미없고 그래픽도 꾸지고** 난장판인 엑스맨을 과거의 이야기로 새로 시작한 메튜 본 감독과 **깔끔하게 다시** 재정리한 브라이언 싱어 감독에게 박수를 **Negative**
CNN-4channel	이렇게 **재미없고 그래픽도 꾸지고** 난장판인 엑스맨을 과거의 이야기로 새로 시작한 메튜 본 감독과 **깔끔하게 다시** 재정리한 브라이언 싱어 감독에게 박수를 **Negative**

emotional expressions, rather than movie-review specific expressions, receive higher attention by the *CNN-Rand*.

Table 14 shows an examples in the same manner as the example illustrated in Table 11. The three models except *CNN-Rand* and *CNN-Static* focus on the negative phrase "재미없고 (boring)" and the positive phrase "깔끔하게 (neatly)". Qualitatively, the former is a stronger emotional expression than the latter, which results in the entire sentence being predicted as negative. However, the CNN-Static finds a stronger positive expression, i.e., "박수를 (pay tribute to)" rather than "깔끔하게 (neatly)", which results in the CNN model predicting the whole sentence as positive.

6. Conclusion

In this paper, we propose CAM^2, a classification and attention model with class activation map, which is a sentiment classification model with word attention based on weakly supervised CNN learning. Although the proposed model is trained based on class labels only, it can not only predict the overall sentiment of a given sentence but also find important emotional words significantly contributing the predicted class. Compared to the previous CNN-based text classification model, CAM^2 utilizes zero-paddings to help the CNN consider every word equally regardless of its position in the sentence. Moreover, it uses average pooling and a large number of filters to preserve the information as much as possible. In addition, various word embedding techniques are employed and integrated.

Experimental results on two movie review datasets, IMDB, which is in English, and WATCHA, which is in Korean, show that the proposed CAM^2 yielded classification accuracies higher than 87% for the IMDB and 78% for the WATCHA dataset. The CNN models that update the word embedding vectors during the sentiment classification learning (*CNN-Rand, CNN-Non-Static, CNN-2channel,* and *CNN-4channel*) achieved higher classification performance than that did not update the word embedding vectors (*CNN-Static*). It is also worth noting that the integration of multiple word embedding techniques improved the classification performance for the IMDB dataset. However, all models showed the ability to find important emotional words in the sentence, although the internal mechanism might be different. For the WATCHA dataset, in particular, the *CNN-Static*, which does not update the word embedding vector during the training, focused more on generally accepted emotional expressions, whereas the other models, which adapt to the language usage pattern in the movie review domain, seemed to focus more on the domain-dependent emotional expressions.

We expect that the proposed methodology can be a useful application in domains where it is important to understand what the input sentences are intended to convey, such as visual question and answering system or chat bots. Although the experimental results were favorable, the current study has some limitations, which lead us to the future research directions. First, the proposed method used a simple space-based token for training word embedding vectors. If more sophisticated preprocessing techniques, such as lemmatization, are performed, the classification and attention performance can be improved. Secondly, quantitative evaluation of word attention, i.e., how good or appropriate the identified words are in the context of sentiment classification, is difficult, which is why we qualitatively interpreted the word attention results in Section 4. Developing a systematic and quantitative evaluation method for word attention can be another meaningful future research topic.

References

Ba, Jimmy, Mnih, Volodymyr, and Kavukcuoglu, Koray. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

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

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Cho, Heeryon, Kim, Songkuk, Lee, Jongseo, and Lee, Jong-Seok. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71:61–71, 2014.

Gui, Lin, Zhou, Yu, Xu, Ruifeng, He, Yulan, and Lu, Qin. Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*, 124:34–45, 2017.

Jo, Eun Kyoung. The Current State of Affairs of the Sentiment Analysis and Case Study Based on Corpus. *The Journal of Linguistics Science*, 61: 259–282, 2012. URL `http://www.dbpia.co.kr/Article/NODE06607901`.

Kalchbrenner, Nal, Grefenstette, Edward, and Blunsom, Phil. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

Kim, Yoon. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Liu, Bing. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Luong, Minh-Thang, Pham, Hieu, and Manning, Christopher D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Maas, Andrew L, Daly, Raymond E, Pham, Peter T, Huang, Dan, Ng, Andrew Y, and Potts, Christopher. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 142–150. Association for Computational Linguistics, 2011.

Medhat, Walaa, Hassan, Ahmed, and Korashy, Hoda. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mnih, Volodymyr, Heess, Nicolas, Graves, Alex, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pp. 2204–2212, 2014.

Nasukawa, Tetsuya and Yi, Jeonghee. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pp. 70–77. ACM, 2003.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, and Sivic, Josef. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694, 2015.

Pang, Bo, Lee, Lillian, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Poria, Soujanya, Cambria, Erik, and Gelbukh, Alexander. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

Ravi, Kumar and Ravi, Vadlamani. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.

Rush, Alexander M, Chopra, Sumit, and Weston, Jason. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Shen, Yatian and Huang, Xuanjing. Attention-based convolutional neural network for semantic relation extraction. In *COLING*, pp. 2526–2536, 2016.

Socher, Richard, Perelygin, Alex, Wu, Jean, Chuang, Jason, Manning, Christopher D, Ng, Andrew, and Potts, Christopher. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Tai, Kai Sheng, Socher, Richard, and Manning, Christopher D. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

Vinyals, Oriol, Kaiser, Łukasz, Koo, Terry, Petrov, Slav, Sutskever, Ilya, and Hinton, Geoffrey. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pp. 2773–2781, 2015.

Xianghua, Fu, Guo, Liu, Yanyan, Guo, and Zhiqiang, Wang. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37:186–195, 2013.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron, Salakhudinov, Ruslan, Zemel, Rich, and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

Yang, Zichao, Yang, Diyi, Dyer, Chris, He, Xiaodong, Smola, Alexander J, and Hovy, Eduard H. Hierarchical attention networks for document classification. In *HLT-NAACL*, pp. 1480–1489, 2016.

Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, Oliva, Aude, and Torralba, Antonio. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.