Opinion mining on book review using CNN-L2-SVM algorithm

M F Roz†, I Muklash†, Soetrisno†, and M Kimura‡

† Mathematics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
‡ Information Science and Engineering, Shibaura Institute of Technology, 3-5-7 Koto-ku Toyosu, Tokyo, Japan

Email: muhammadfakhrurrozi95@gmail.com, imammukhlass@gmail.com, soetrisno@matematika.its.ac.id, masaomi@shibaura-it.ac.jp

Abstract. Review of a product can represent quality of a product itself. An extraction to that review can be used to know sentiment of that opinion. Process to extract useful information of user review is called Opinion Mining. Review extraction model that is enhancing nowadays is Deep Learning model. This Model has been used by many researchers to obtain excellent performance on Natural Language Processing. In this research, one of deep learning model, Convolutional Neural Network (CNN) is used for feature extraction and L2 Support Vector Machine (SVM) as classifier. These methods are implemented to know the sentiment of book review data. The result of this method shows state-of-the art performance in 83.23% for training phase and 64.6% for testing phase.

1. Introduction

Recently, internet users in this world are increasing every year. It makes internet become large marketplace for people to buy or sell their product on internet. They use electronics and technology infrastructure to make transaction (selling, buying, transfer, or exchanging product, service, and/or information), included in business, business to business interaction, and costumer to business, this action is called E-commerce [1]. However, there are still many problems and enhancements for the implementation that makes many comments, suggestions, and critics emergence in purchasing of product. Review of a product defines the quality of product itself that has been used. That’s review can be the opinions that represent good or bad product. But it is so hard to interpret real meaning of that opinion in the implementation. It shows us that it is necessary to extract the meaning of review in order to know the sentiment of that review. The process to extract useful information from user review is called opinion mining. This process is applied by using natural language processing and text analysis methods [2]. A review or opinion is processed in order to know the emotion that is written in that opinion accurately, so that we can get the information of quality of a product that can be classified become good, neutral, or bad. Review of a product has an important role in business. Customer will trust to a product that have good review from other customer. This makes many sellers need positive review for their product in order to be sold well.

There are many methods that have been used in opinion mining. The research from Pang and Lee tried to mine the opinion of movie review by using Naïve Bayes and Support Vector Machine (SVM), besides that they used Based on Minimum Cut for feature selection. That research yielded 86.4% accuracy [3]. Besides that, there is a model which is more modern and enhancing nowadays.
called Deep Learning Model. One of method that implement Deep Learning model is Convolutional Neural Network (CNN). Generally, CNN is used for image processing as classifier or cluster. Meanwhile, Tang proposes in his research, combination of CNN and L2-Support Vector Machine (L2-SVM) that is tested on image data [4]. His research tried to change activation function of CNN that usually use softmax function become L2-SVM. After that, in 2014 Kim delivered his innovation about implementation of CNN on sentence classification that shown state-of-the art performance [5].

Therefore, this research use combination of CNN and L2-SVM methods for opinion classification. This method will yield two kind of class of opinions, positive and negative. CNN method is used likes Kim’s research but activation function is L2-SVM likes Tang’s research. Testing of this method is used in book review data that is got from website amazon.com. Besides that, performance of method is measured as consideration for next research.

2. Preliminaries

2.1. Related works
There are many methods that had been used by researcher for mining an opinion. It started from a research from Pang and Lee whose were tried to mine opinion from movie review by using Naïve Bayes and Support Vector Machine (SVM) also by using feature selection, Based on Minimum Cut. Their research concluded that their method could reach 86.4% performance [3]. Next, research about opinion mining continued with many kinds of method and combination such as Maximum Entropy, Backpropagation, K-means, that yielded performance 85.4%, 86%, and 78% respectively [3]. Besides that, there is a model which is more modern and enhancing nowadays called Deep Learning Model.

In 2013, Tang introduced his research about implementation of deep learning which is combined with SVM. One of deep learning method that’s used by him is Convolutional Neural Network (CNN). In that research, He tried to compare performance between deep learning generally and combination with SVM. He changed activation function of CNN that usually use softmax function become L2-SVM. It was done because Tang thought that there is similarity between softmax classifier and L2-SVM classifier. The result is error of L2-SVM is less than softmax. Because of that, He concluded that changing softmax become L2-SVM can easily help classification process [4].

After that, in 2014, Kim introduced implementation of CNN method on a sentence classification. That research proposed deep learning model that’s applied in Natural Language Processing (NLP). Kim made conclusion in his research that simple CNN with one convolution layer can yield state-of-the art performance.

![Figure 1. Model architecture with two channel for a sample sentence](image)

2.2. Convolutional neural network
Convolutional Neural Network (CNN) is advanced of Multilayer Perceptron (MLP) that is designed to process two-dimension data. CNN belongs to deep neural network because the depth of network is high and can be applied on many image data. Mechanism of CNN has similarity with MLP, but every neuron in CNN can be represented in two-dimension form that is not likes MLP which every neuron is one-dimension form.
The architecture of CNN consists of two main layers, convolution layer and pooling layer. Implementation of CNN on sentence classification is enough to use two main layers. It can be illustrated on figure 1.

2.2.1. Convolution layer
Convolution layer is first layer in CNN-L2-SVM network architecture. In this layer, convolution process is done to all word vectors of review. Figure 2 shows the example of convolution process on a sentence.

Let the length of a review sentence is $n$, therefore there are $n$ word vectors with size of dimension is 50 as the input of model. Assume $x_i \in \mathbb{R}^{50}$ is the $i$-th word vectors of a sentences. If word vectors are concatenated, the review will be able to be represented like below:

$$x_{1:n} = x_1 \oplus x_2 \oplus x_3 \oplus ... \oplus x_n$$

(1)

Where $\oplus$ is concatenation operator. Generally, $x_{i:i+j}$ means the result of word vectors concatenation form $i$-th indices until $i + j$-th.

$$x_i, x_{i+1}, x_{i+2}, ..., x_{i+j}$$

(2)

Convolution operation us a filter or parameter $w \in \mathbb{R}^{50h}$ where $h$ is the size of window. In order to get feature value, we can get equation likes below:

$$c_i = f(\text{net})$$

(3)

$$\text{net} = w \cdot x_{i:i+h-1} + b$$

(4)

At that equation, there is bias parameter $b \in \mathbb{R}$, $c_i \in \mathbb{R}$, are feature values that is yielded in $i$-th indices, and $f$ is non-linear function. In this research, non-linear function that is used is Rectified Linear Unit (ReLU) function. Output of that function gives output constraint with positive value. The function can be written as below:

$$\text{ReLU}(x) = \max(0,x)$$

(5)

So that equations can be represented by:

$$c_i = \text{ReLU}(\text{net})$$

(6)

$$\text{net} = w \cdot x_{i:i+h-1} + b$$

(7)

Filter $w$ can be applied to each word window possible in sentence of review $\{x_{1:h}, x_{2:h}, x_{3:h}, ..., x_{n-h+1:n}\}$. As the result, we can get feature map likes below:
\[ c = [c_1, c_2, c_3, ..., x_{n-h+1}] \] (8)

2.2.2. Pooling layer
After we have got feature map that comes from convolution layer, so each of feature map can be processed in pooling layer. In this layer, the most importance value of each feature map is taken. Mathematically pooling operation can be written like below:

\[ \hat{c} = \max\{c\} \] (9)

Where \( \hat{c} \) is maximum value of feature map \( c \) that is correlated with a filter. Because there are \( m \) filter, so the result of pooling layer is a vector that consists of maximum values of each feature map and there are \( m \) element. Therefore, we can obtain a vector below:

\[ z = [\hat{c}_1, \hat{c}_2, \hat{c}_3, ..., \hat{c}_m] \] (10)

Where \( z \) is output vector from pooling layer that will input into the next layer.

2.3. L2-support vector machine
Linear support vector machine is firstly formulated for binary classification. Let training data and connected label is \((x_n, y_n), n = 1, ..., N, x_n \in \{-1, +1\}\), SVM learning consists of constraint optimization likes below:

\[
\begin{align*}
\min_{w, \xi_n} & \quad \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\
\text{s.t.} & \quad w^T x_n t_n \geq 1 - \xi_n \quad \forall n \\
& \quad \xi_n \geq 0 \quad \forall n
\end{align*}
\] (11)

\( \xi_n \) is slack variable which penalizes data point which violate margin requirements. Note that we can include the bias by augment all data vectors \( x_n \) with a scalar value of 1. The corresponding unconstrained optimization problem is the following:

\[
\min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n t_n, 0)
\] (12)

Objective equation (12) is known as main constraint of L1-SVM, with standard hinge loss. Because L1-SVM is not differentiated, so that the most popular version is known as L2-SVM which minimalize squared loss.

\[
\min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n t_n, 0)^2
\] (13)

L2-SVM can be differentiated and cause more loss for pints which violate margin. To predict class label of testing data \( x \):

\[
\arg\max_t (w^T x) t
\] (14)
2.4. Algorithm

This research uses CNN and L2-SVM algorithm in opinion mining on review. CNN algorithm consists of 3 main layer, convolution, pooling, and fully connected layer. The input of this algorithm is a bunch of word vectors that is obtained from pre trained process. That CNN algorithm can be explained in pseudocode below:

```
input : a data set of opinion sentences that is already represented by word vectors
output : class label of each data
window_size = \{h_1, h_2, h_3, ..., h_n\},
z = \emptyset
data = \{X_1, X_2, ..., X_m\}
for each h in window_size:
    w ← initializeFilter(h, m)
    for X in data:
        s ← size(X)
        c = \emptyset
        for i in 0:s - n + 1
            x ← concatenate(X_{i+d_h-1})
            temp ← nonLinear(w^T x + b)
            c ← c ∪ temp
        end
        ̂c ← max(c)
        z ← z ∪ ̂c
        probability ← softmax(z)
    end
end
```

Since there is a little bit modification of CNN in this research so that the last layer, fully connected layer is changed by SVM. The result of these algorithms is not a probability but score of class which determine label class of data. Then, L2-SVM loss function is used in training phase that influence the back propagation process to get optimal hyper parameter. The pseudocode of the combination between CNN and L2-SVM is written below:

```
input : a data set of opinion sentences that is already represented by word vectors
output : class label of each data
window_size = \{h_1, h_2, h_3, ..., h_n\},
z = \emptyset
data = \{X_1, X_2, ..., X_m\}
for each h in window_size:
    w ← initializeFilter(h, m)
    for X in data:
        s ← size(X)
        c = \emptyset
        for i in 0:s - n + 1
            x ← concatenate(X_{i+d_h-1})
            temp ← nonLinear(w^T x + b)
        end
        \hat{c} ← \max(c)
        z ← z ∪ \hat{c}
        probability ← softmax(z)
```
3. Result and discussion

3.1. Result

Dataset that is used for this research is book review dataset from Amazon.com that contain positive and negative review. There are 8 kind of books that have different distribution of positive and negative review. We can see from figure 3 the distribution of each books. Based on figure 3, we can see that the best distribution among other books is “Gone Girl”. Proportion of positive review is less than others, so that the model that we will build is good enough to predict class label of book review.

After through training phase, we can get model which fit input data. The model is required to be tested the accuracy so that we can conclude that model is suitable model or not. Testing or validation phase is done in order to test data that has been separated before independently. Testing and training data is independent data so that we can see the ability of model to overcome new data. According to implementation result that had been done, model yield the accuracy of each iteration for all kind of data. On figure 4 shows the accuracy of model during training phase that had been done 20 times iteration.

From figure 4 we can see that the accuracy of model with training data is very good. Just 20 times iteration, the accuracy can reach 83.23% and will increase for next iteration. On the other hand, in order to get that result, it takes long times around 76 minutes.

Meanwhile, the accuracy of model with testing data had lower result than training data. It happens because there are not supervising of the result for classification in testing data. figure 5 illustrates the accuracy of training data in 20 times iteration. We can see figure 5 above that accuracy of testing data is fluctuated if we compare with the result of training data. The accuracy of model with testing data yield final result 61.94%. We can see from figure 5 that the accuracy of testing data decrease from 66% until 60 – 61% at the last iteration. If we see the result of testing and training phase that the graph is in contrary. It means that model has overfitting. Overfitting is an event where model is too good in training data but too bad in testing data. This problem is common in classification with artificial neural network.

3.2. Discussion

Model that has been created is must evaluated. The model need to know the performance for doing specific task. There are several evaluation methods that can be done for classification on a model such as accuracy, precision, recall, and F1 score. In this research, there are four evaluation method are used to know the performance of model. We can see the result in figure 6.
Figure 3. Distribution of book review label

Figure 4. Graph of training data accuracy movement

Figure 5. Graph of testing data accuracy movement
According to that graph, we can see that the accuracy of testing data is around 0.646 or 64.6%. For the precision, we got 0.646 or 64.6% too. Meanwhile, recall value is around 0.647 or 64.7%. That is little bit different with F1 score. We got value around 0.644 or 64.4% which is lower than others. These result shows us that model have good enough performance to classify book review data with accuracy, precision, recall, and F1 score are in same value around 64%.

4. Conclusion
According to the result, it can be concluded that model conduction of Convolutional Neural Network (CNN) and L2 Support Vector Machine (L2-SVM) are able to determine sentiment or opinion of book review. Besides that, we get the performance of model to determine sentiment and opinion of review around 83.23% for training phase and 64.4% for testing phase.

5. References
[1] Manzoor, Amir. 2010. E-Commerce: An Introduction. Saarbrucken: Lap Lambert
[2] Ali, Farman.Kwak, Kyung-Sup dan Kim, Yong-Gi. 2016 Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. Elsevier. 47:235-250
[3] Singh, V. and Dubey, S. K. Opinion Mining and Analysis: A Literature Review. IEEE 5th International Conference- Confluence The Next Generation Information Technology Summit. 2014. page 236-237
[4] Tang, Y. 2013 Deep Learning using Linear Support Vector Machines International Conference on Machine Learning 2013: Challenges in Representation Learning Workshop. Atlanta USA.
[5] Kim, Y. 2014 Convolutional Neural Networks for Sentence Classification Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). page 1746-1751