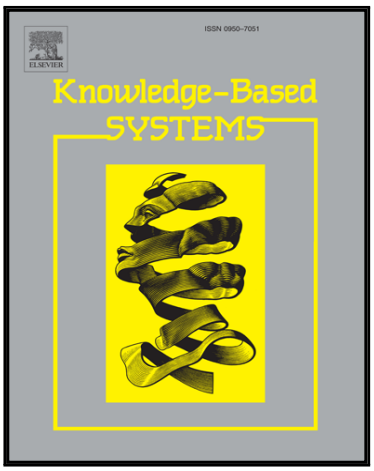

Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network

Gichag Lee , Jaey un Jeong , Seungwan Seo , CzangYeob Kim , Pilsung Kan g

PII: S0950-7051(18)30171-0
DOI: 10.1016/j.knosys.2018.04.006
Reference: KNOSYS 4288

To appear in: *Knowledge-Based Systems*

Received date: 18 September 2017
Revised date: 2 April 2018
Accepted date: 3 April 2018

Please cite this article as: Gichag Lee , Jaey un Jeong , Seungwan Seo , CzangYeob Kim , Pilsung Kan g , Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.04.006

Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network

Gichag Lee, Jaeyun Jeong, Seungwan Seo, CzangYeob Kim, Pilsung Kang*

School of Industrial Management Engineering, Korea University, Seoul, South Korea

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 a lack of 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. To verify the proposed methodology, we evaluated the classification accuracy and the rate of polarity words among high scoring words using two movie review datasets. Experimental results 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 Localization, Convolutional Neural Network, Class Activation Mapping, Sentiment Analysis

* Corresponding author: pilsung_kang@korea.ac.kr