Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 111 (2017) 376–381

Procedia
Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Information Technology, IAIT2016, 19-22 December 2016, Macau, China

# CNN for situations understanding based on sentiment analysis of twitter data

Shiyang Liao[a, b]*, Junbo Wang[b], Ruiyun Yu[a], Koichi Sato[b], Zixue Cheng[b]

[a]Northeastern University, NO. 3-11, Wenhua Road, Heping District, Shenyang, 110819, China
[b]University of Aizu, Kamiiawase 90, Ikkimachi Tsutuga, Aizu-wakamatsu, 965-8580, Japan

## Abstract

In this paper, we propose an approach to understand situations in the real world with the sentiment analysis of Twitter data base on deep learning techniques. With the proposed method, it is possible to predict user satisfaction of a product, happiness with some particular environment or destroy situation after disasters. Recently, deep learning is able to solve problems in computer vision or voice recognition, and convolutional neural network (CNN) works good for image analysis and image classification. The biggest reason to adopt CNN in image analysis and classification is due to CNN can extract an area of features from global information, and it is able to consider the relationship among these features. The above solution can achieve a higher accuracy in analysis and classification. For natural language processing, texts data features also can be extracted piece by piece and to consider the relationship among these features, but without the consideration of context or whole sentence, the sentiment might be understood wrong. And currently, convolutional neural network is one of the most effective methods to do image classification, CNN has a convolutional layer to extract information by a larger piece of text, so we work for sentiment analysis with convolutional neural network, and we design a simple convolutional neural network model and test it on benchmark, the result shows that it achieves better accuracy performance in twitter sentiment classification than some of traditional method such as the SVM and Naive Bayes methods.

*Keywords:* machine learning, deep learning, convolutional neural network, sentiment analysis, data mining

* Corresponding author. Tel.: +81-242-37-2712; fax: +81-242-37-2732.
*E-mail address:* liaosy@swc.neu.edu.cn; d8172103@u-aizu.ac.jp

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the organizing committee of the 8th International Conference on Advances in Information Technology
10.1016/j.procs.2017.06.037

## 1. Introduction

Social media has become a source of varied kind of information., and the new type information could be harvested from social media. As one of the most popular social media, Twitter has at least 100 million active users, furthermore, 572,000 new accounts has been created on a single day (March 12, 2011, the day after the Sendai earthquake and resulting nuclear disaster), while an average of 140 million tweets are sent daily[14]. Valuable knowledge is often hidden behind Twitter contents and cannot be easily processed through automation[12]. Twitter is an ideal social media for the extraction of general public opinion on specific issues[7]. Twitter data is useful for sentiment analysis, such as opinion mining or natural language processing[10].

There are several approaches for sentiment analysis on Twitter, one of them is machine learning. Deep learning models have achieved great results in computer vision[6] and speech recognition[3] in recent years. To solve NLP (Natural Language Processing) problems, machine learning is also useful by using a general learning algorithm combined with a large sample of data to learn the classification rules. Several methods do it with traditional algorithm such as SVM or Naïve Bayes, most of such methods consider text word by word, classify a sentence to positive or negative by analyzing the word in the text, sometimes information lost by extracting key word without other word; CNN (Convolutional Neural Networks) is one of machine learning models which has archived impressive result in image recognition several years ago and has archived remarkable results in natural language processing recently, there is a convolutional layer to make a piece of words can be considered together. In this paper, we propose an approach to parsing Twitter data to understand situation in the real world based on a CNN model to do the sentiment analysis. We adopt convolutional neural network as our sentiment analysis model because in image analysis and classification field, CNN can extract an area of features from global information, with the convolution operation, a piece of data information can be extract together as the features, and it is able to consider the relationship among these features. For computer vision, such as image analysis, it is able to extract a part of pixel data information, not only extract the pixels one by one, the features information can be extracted piece by piece, the piece contains multi pixels data information; when we transfer the text into matrix, it can also be considered as same as an image pixels' matrix, so we can do the same operation to the text data to make the input features to the model can be trained in another effective way.

The paper is organized as follow. In section 2 we introduce the approach we propose, and the key model we have choose in this approach; section 3 describe the experiment data, experiment method and experiment results, also include discussion about the experiments and the results; section 4 is the conclusion of this paper and future plan.

## 2. Approach

The structure of this approach is shown in Figure 1. There is a convolutional neural network on the right side of the figure, which will be introduced in detail at the next sub section of this paper. Our approach is based on machine learning approach. A sentiment can be simply categorized into two groups[13], so we choose MR[1] and STS Gold Dataset[16] which has been labelled into two groups, positive and negative, as training dataset[9]. Here the MR is a set of movie reviews with one sentence per review, the reviews are from Internet users and are similar to Twitter data; the STS Gold Dataset is an essential collection of real Twitter dataset. After training of the CNN with dataset, we input Twitter data we have gotten by hashtag and stored in MongoDB. The convolutional neural network will output the sentiment. With the sentiment and the geo-tag or other information included in Twitter, we can do further research.

### *2.1. CNN Model*

The example of CNN model is shown in Figure 2, the model is simplified convolutional network of the CNN for sentence classification[5], but we use single channel for simpler. Firstly, we need to transfer the sentence into matrix,

[1] https://www.cs.cornell.edu/people/pabo/movie-review-data/

the rows of each sentence matrix are word vector representations. The dimensionality of the word vectors is $d$. When the length of a sentence is $s$, the dimensionality of the sentence matrix is describes as $s \times d$ in[5, 15]. According to Collobert and Weston[2], we consider the sentence matrix as same as an image matrix, we perform convolution on the matrix with linear filters. The height of the filter is the region size of the filter.

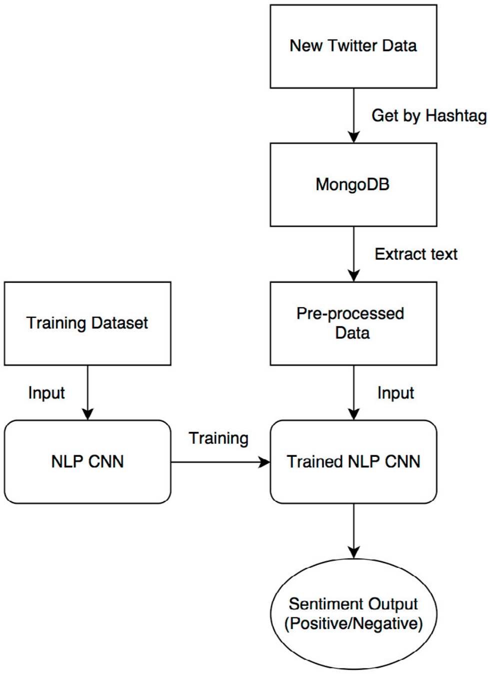

Fig. 1. Proposed approach.

There is a filter which parameterized by the weight matrix w with region size $h$; the sentence matrix $A \in \mathbb{R}^{s \times d}$, $A[i:j]$ is the sub-matrix of $A$ from row $i$ to row $j$. The output $o \in \mathbb{R}^{s-h+1}$ can be calculated

$$A = \pi r^2 o_i = w \cdot A[i:i+h-1] \tag{1}$$

In the formula, $i = 1 \ldots s - h + 1$ is the dot product between sub-matrix and the filter. Same as other neural network model, bias $b \in \mathbb{R}$ and activation function $f$ have been added to $o_i$, and the feature map $c \in \mathbb{R}^{s-h+1}$ for:

$$c_i = f(o_i + b) \tag{2}$$

After that, we apply a pooling function to each feature map to get a fixed length vector, we use 1-max pooling[1] to extract from feature map. Then with a softmax function to get the final classification. At the softmax layer, we apply dropout as a means of regularization[4]. Also we perform an l2 norm constraint, which is effective for overfitting when training of the neural network.

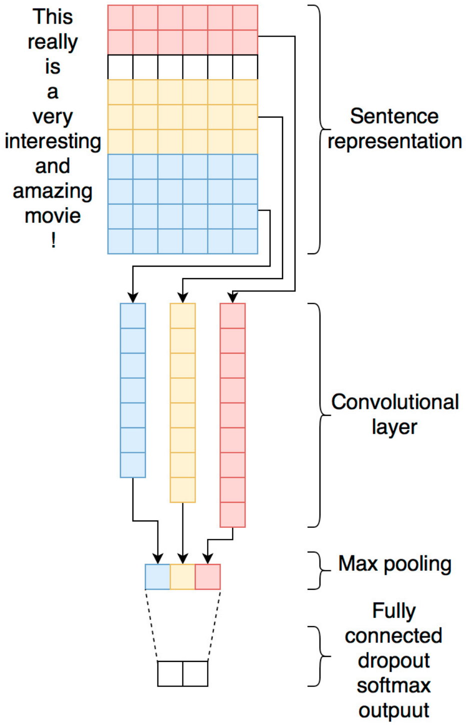

Fig. 2. CNN model architecture for an example.

## 3. Experiment and discussion

We choose MR and STS Gold Dataset as benchmark dataset[9, 16], MR is a collection of movie review and STS Gold dataset is a collection of real tweets, the detail of these two dataset is shown in Table 1. We split each of the dataset into 2 groups, train set which to be input as training samples and the development set to verify the accuracy of the checkpoint of the convolutional neural network; for each of the dataset, we set the train set around 90% of the whole data amount and the development set is around 10%. For the testing, we train the convolutional neural network model several times, the one with highest average development accuracy one we set filter windows (h) 4, 5, 6 with 100 feature maps each, dropout rate (p) is 0.5, l2 regularization lambda is 0.001, the batch size is 64. Training with the values above give a results is shown in Figure 3 and Figure 4.

Table 1. Detail of MR and Gold Dataset.

| Dataset | No. of sentences | Positive | Negative |
|---|---|---|---|
| MR | 10662 | 5331 | 5331 |
| STS Gold | 2034 | 632 | 1402 |

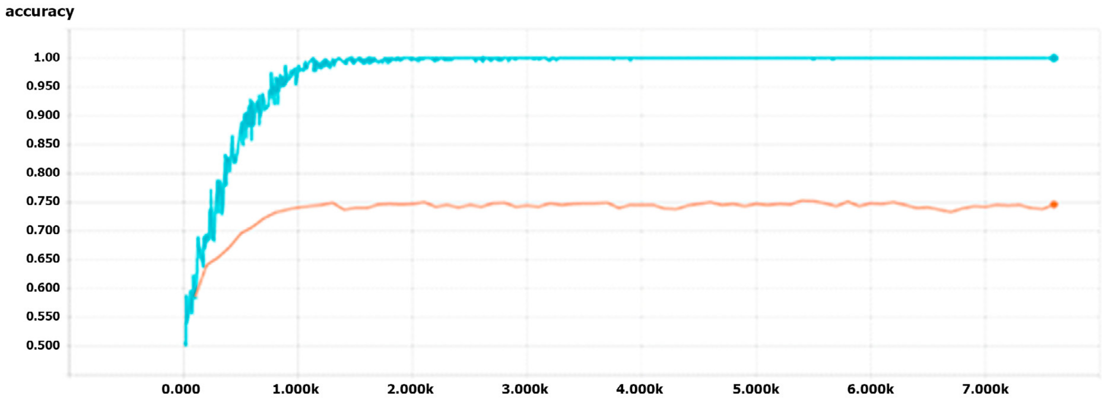

Fig. 3. MR dataset training results.

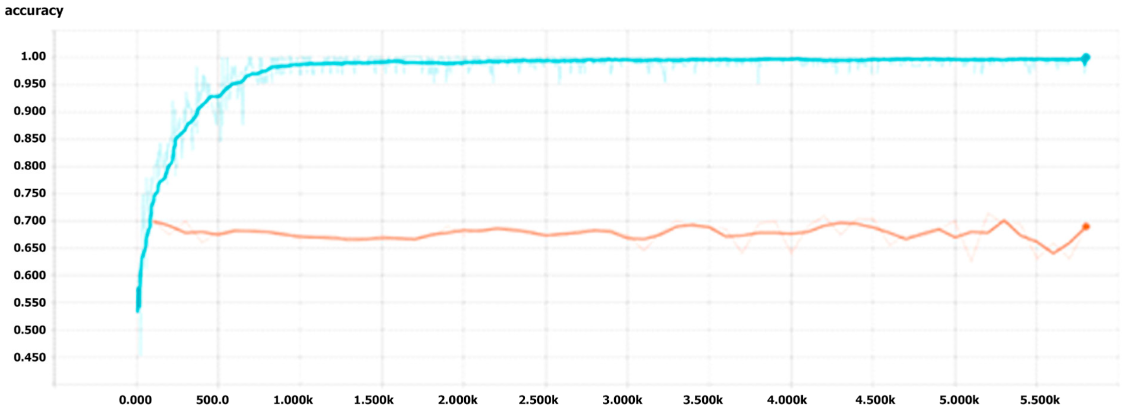

Fig. 4. STS Gold dataset training results.

As shown in Figure 3, the cyan is train set accuracy, and the orange is development set accuracy, the development accuracy of MR training results is 74.5%, the development accuracy of STS Gold Dataset training results is 68%, and as the cyan line shown, during the training of the network, there are lots of overfitting though l2 norm constraint has been applied. Ref.[17] to do the twitter sentiment analysis uses a data-set of 1709 instances for each class, the dataset is also in 2-way, positive and negative, and get an average accuracy of 75.39% which is near to our result of MR training. However, the proposed approach to analysis twitter data is a demonstration to use CNN to classify the tweets. In the future, the model can be redesigned as a deeper convolutional neural network, and train with larger dataset. Also, for our proposed approach, we can use the trained convolutional neural network to understand the situation in the real word with Twitter data sentiment analysis, such as the predicting of user satisfaction of a product, feeling with some particular environment or destroy situation after disasters, or work on other research with other information included in Twitter data, such as spatial data of geo-tag or multimedia data.

## 4. Conclusion

In present work we have proposed an approach for Twitter data sentiment analysis, and in this approach, we construct a convolutional neural network for sentiment analysis based on text. There many further research can be done, such as considering the word2vec tool, multilayer convolutional neural network, larger training dataset and other situation or status analysis.

## Acknowledgements

This research was supported by the Fundamental Research Funds for the Central Universities under Grant no. N130817003 and JST, Strategic International Collaborative Research Program, SICORP, entitled "Dynamic Evolution of Smartphone-Based Emergency Communications Network", from 2015 to 2018.

## References

1. Boureau Y, Ponce J, LeCun Y, A theoretical analysis of feature pooling in visual recognition, In Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, p. 111–118.
2. Collobert R, Weston J, A unified architecture for natural language processing: Deep neural networks with multitask learning, In Proceedings of the 25th international conference on Machine learning, 2008, p. 160–167.
3. Graves A, Mohamed A, Hinton G, Speech recognition with deep recurrent neural networks, In Proceedings of ICASSP 2013, 2013.
4. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR, Improving neural networks by preventing coadaptation of feature detectors, arXiv preprint, 2012, arXiv:1207.0580.
5. Kim Y, Convolutional Neural Networks for Sentence Classification, arXiv preprint, 2014, arXiv: 1408.5882.
6. Krizhevsky A, Sutskever I, Hinton G, ImageNet Classification with Deep Convolutional Neural Networks, In Proceedings of NIPS 2012, 2012.
7. Osimo D, and Mureddu F, Research Challenge on Opinion Mining and Sentiment Analysis, In Proceeding of the 12th conference of Fruct association, United Kingdom, 2010.
8. Pak A, Paroubek P, Twitter as a corpus for sentiment analysis and opinion mining, In Proceedings of LREC, 2010.
9. Pang B, Lee L, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, In Proceedings of ACL, 2005, p. 115-124.
10. Rambocas M, Gama J, Marketing Research: The Role of Sentiment Analysis. In Proceeding of The 5th SNA-KDD Workshop'11. University of Porto, 2013.
11. Read J, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, In ACL. The Association for Computer Linguistics, 2005.
12. Sahito F, Latif A, Slany W, Weaving twitter stream into linked data a proof of concept framework, In 7th International Conference on Emerging Technologies (ICET), 2011, p. 1-6.
13. Saif H, He Y, Alani H, Semantic Sentiment Analysis of Twitter, Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data, United Kingdom: Knowledge Media Institute, 2011.
14. Vatsavai RR, Chandola V, Klasky S, Ganguly A, Stefanidis A, Shekhar S, Spatiotemporal Data Mining in the Era of Big Spatial Data: Algorithms and Applications, BigSpatial '12 Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, ACM, 2012, p. 1-10.
15. Zhang Y, Wallace BC, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, arXiv preprint, 2016, arXiv: 1510.03820v4.
16. Saif H, Fernandez M, He Y, Alani H, Evaluation Datasets for Twitter Sentiment Analysis, A survey and a new dataset, the STS-Gold, Workshop: Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM) at AI*IA Conference, Turin, Italy, 2013.
17. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R, Sentiment Analysis of Twitter Data, In Proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media, 2011, p. 30-38