Convolution Neural Network for Text Mining and Natural Language Processing

N I Widiastuti
Department of Informatics Engineering, Universitas Komputer Indonesia, Indonesia

E-mail : nelly.indriani@email.unikom.ac.id

Abstract. The objective of this study is to get an overview of the improvements applied in a number of studies and problems that have not been resolved. We have surveyed more than 30 scientific articles obtained from scientific article portals such as Science Direct, IEEE explore, Arxiv, and Google Scholar. Based on this abstract, we obtain similarities and differences based on the problem solved, the preprocessing method for data input, and the approach taken to achieve the goal. The results show that some problems have not been resolved by CNN in the text mining domain and NLP. This happens because CNN has been used to solve problems in each case such as sentiment analysis, classification of documents or NLP cases such as entities and their relationships, or semantic representation. CNN has shown to be very effective in areas such as image recognition or classification. However, at present, there are many CNN studies that use text as processed data. Text mining and Natural Language Processing (hereinafter referred to as NLP), is the most interesting research domain for now. Various development of sentiment analysis, information extraction, document classification, or the introduction of entities extensively entering business subjects, medicine, or data security.

1. Introduction
Convolutional Neural Networks, known as CNN, is a category of Neural Networks that uses a multilayer perceptron variation that is designed for minimal preprocessing. CNN is also known as shift invariant or space invariant artificial neural networks (SIANN), based on shared-weights architecture and invariance translation characteristics[1]. CNN has proven to be very effective in areas such as image recognition or classification. However, at present, there are many CNN studies that use text as processed data. Text mining and Natural Language Processing (hereinafter referred to as NLP), is the most interesting research domain for now. Various development of sentiment analysis, information extraction, document classification, or the introduction of entities extensively entering business subjects, medicine, or data security.

Some research has shown the success of CNN in processing images. The research claims that CNN has succeeded in identifying traffic signs [2], the robot that can recognize objects [3] and is able to classify images into 1000 classes [4]. However, it does not mean that CNN is only able to work for image data; some studies have also tested text data. A literature review has been conducted on CNN [5] for image, sound, and text data. In the survey paper, produces a taxonomy that describes the structure of CNN that has been done. The taxonomy can be seen in Figure 1.
In Figure 1, a detailed structure of CNN can be shown in the surveyed article, but it does not specifically address cases of text mining and NLP domains. Other surveys have also been conducted on the influence of input vectors, filter sizes, a number of map features, activation functions, pooling strategies, and regularizations in CNN to classify documents [6]. In the article provide some suggestions for further research. In our previous research, a survey of deep learning has been conducted in the text-mining and NLP domains [7]. In the survey, deep learning is generally discussed in the same domain. Nonetheless, the survey did not specifically explain the use of CNN.

More than 30 studies were discussed in this survey study. The discussion focuses more on the cases that have been resolved by CNN, how preprocessing is done, what is the function of CNN in solving problems in the research, and how the CNN architecture is used. The purpose of this study is to provide an overview of research results in CNN as a method used in text mining and NLP.

2. Methods
A survey paper aims to convey similarities and differences (gaps) that appear in the reference paper. To achieve this in this study, we collected articles from several credible sources such as Elsevier, google scholar, IEEE, and Science Direct. All articles collected were related to CNN in the case of text mining and NLP domains. The article has been declared accepted or is in the process of being reviewed in the 2015-2018 period. Once collected, all articles are summarized to obtain the core of each research (problem, preprocessing, CNN network architecture, and CNN functions). Based on the summaries obtained, we group those that have similarities and separate those that have differences. At the end of this method, discuss what can still be done by other studies that will use CNN to solve cases with text data.

3. Results and Discussion
In this section the discussion begins with the equations found in the articles collected, then the differences in the problems solved, the input structure, the CNN function in solving problems, and the CNN architecture used.

3.1 Applications
CNN application in the case of data security found 5 problems regarding spam, code injection, and malware detection. Detection of code injection on a mobile device was carried out by Ruibo Yan et al. [8]. The test results on the application obtain precision of 98.48%. In the case of malware detection, 2 studies were found, namely those carried out by ElMouatez Billah et al [9] dan Fabio Martinelli et al[10]. They use CNN to detect malware on Android mobile devices. Although the data they use is not the same, both from the source and the amount, the results obtained are very satisfying, which is around
85% to 95% and 96% -99%. For spam detection problems, Luyang Li et al and Yafeng Ren et al. Have different research questions. Luyang Li et al. Focused more on research on presenting data and looking for combinations of features that were considered optimal[11]. In the test, the results obtained that CNN and its variations are no better than the proposed method. Whereas Yafeng Ren et al. Used CNN to detect spam opinions with unsatisfactory results compared to Bidirectional Average Gated Recurrent Neural Network (GRNN)[12].

In the case of analytical sentiment, many studies have been carried out using CNN. Sentiment analysis in the aspect level or done by Souja Poria et al using Deep CNN plus Linguistics Patterns for sentiment aspects in product reviews obtained 87% accuracy[13]. In another study Souja Poria et al. Also classified sentiments based on visual, audio and textual[14]. Overview of aspects also develops not only on objects but also on external factors that influence sentiment[15] and add emoji elements as features[16]. Besides focusing on cases, there is also research on sentiments that test parameters[17], network[18], and classification technique[19].

Research in classifying and summarizing documents is quite dominating. 15 studies were found related to this, and 7 of them used medical data. Medical cases raised are quite diverse, ranging from breast cancer[20], medical records[21-23], online medical consultation [24], linkages of drugs to diseases[25], and mental health[26]. Another focus is to use CNN to see the relationship between one entity and another, such as cause-effect, component-overall, employee-leader, etc [27]. In several articles found, there are studies that try to link the image with the title. Recognition is done at the character level [28-29]. Not only classifications of sentiments or documents, but CNN is also used for summarizing singles and multi-documents [30-31].

Natural language processing or NLP is an important element in text mining. In this study, there were 7 studies using CNN to obtain a linguistic analysis. To detect paraphrasing in a document, Basant Agarwal et al. Combine CNN with Recurrent Neural Network or RNN[32]. The recognition of the entity and its relationship is carried out by Suncong Zheng et al. by combining CNN with Bidirectional Long Short Term Memory [33]. Jiaming Xu et al tried to produce semantic features in the classification of short documents[34]. The automatic annotation system was created by Baiwei et al on sentiment by combining CNN with BiLSTM [35].

3.2 Feature Representation
How to represent data or in this case in the form of text or documents has a clear influence on the results to be achieved. Ye Feng and Byron suggested using word vectors instead of one hot encoding[6]. However, Jiaming Xu and colleagues compiled words into binary codes using unsupervised dimensionality reduction methods. The results obtained are claimed to be successful by combining learning representations[34]. Zufan Zhang et al. Used Semantic embeddings, sentiment embeddings, and lexicon embeddings to gain attention in the case of analytical sentiment. The results obtained are quite promising because these representations are considered to represent global features[36].

3.3 Network Architecture
This section is used to explain the various types of network architecture. They use the Hybrid Deep Learning Network, a combination of CNN, Long Short Term Memory (LSTM), and Fully Connected Layer[8]. Tests carried out on several CNN, and SCNN architectures cannot be said to be superior when compared to Sentence Weight Neural Network (SWNN) in the case of detection of review spam[11], and various network combinations in the case of analytical sentiment[18]. The hybrid neural network is bidirectional encoder-decoder LSTM module (BiLSTM-ED) for entities extraction. CNN module for classification of relations [33]. Complete information generated by extraction uses a new pooling scheme and a combination of convolutional layer and the bidirectional long-short-term memory [37]. Three different attentions including attention vectors, LSTM (Long Short Term Memory) attention and attentive pooling are integrated into CNN models [36]. A tag prediction model based on convolutional neural networks (CNN) and bi-directional long short term memory (BiLSTM) networks [35]. For optimal clusters by employing K-means to cluster the learned representations [34].
In a study conducted by Zufan Zhang et al, CNN was used as part of the feature extraction process. 3 CNN networks are used to process 3 types of input representations to get attention. The resulting output then uses the regression concept proposed to classify short documents[36].

Based on the survey results that we conducted in this article, various developments carried out by the researchers have been discussed thoroughly. A number of studies conducted tests on CNN in various cases. In text mining and NLP domains, CNN has proven its ability to overcome the problem of text or natural language data. Even when CNN compared to RNN to classify medical records documents, there was a little anomaly. The basic character of CNN is designed to process data in the form of matrices, not for data sequences, but CNN can outperform RNN [26]. However, in some studies, CNN is still combined with RNN, LSTM or BiLSTM [32][33][35].

Representative features should have a strong influence on accuracy. There are not many studies that measure the influence of various feature representation on the performance of CNN in text mining and NLP domains. Our survey results show that representation in the form of word embedding is more trusted as a form of input than one hot encoding. But not so with research that represents words to be more flexible in obtaining semantic features [34]. In this viewpoint, CNN can be used as a method for extracting features [32][15][12][29] or model sentences [37]. The resulting value is then used by RNN, BiLSTM [33], SVM, K means, etc.

The working mechanism or method proposed in a study must be unique. Research contributions are seen from the differences between research with one another in the same domain. The neural network architecture is one part of the mechanism. In this survey, we found that CNN is not only the main algorithm for answering research questions, but also as an algorithm that supports the functions of other algorithms. Based on our survey, several studies combined CNN with other neural networks [32-33], regression[24] or clustering[27][34]. This one-sided merger has positive expectations, but cannot be denied undesirable consequences. All research aims to get high accuracy. Meanwhile, high computing will haunt the application of multiple layers of algorithms. This will be felt in a lot of data usage. Besides combining, CNN also experiences variations in the structure and number of convolution layers[21][8].

4. Conclusion
The survey results that we conducted have produced reviews from various perspectives. In general, further research suggested is more to test in other cases. Based on this, the first suggestion we can convey is to test the computational efficiency of the CNN architecture. This is important to obtain test results not only on accuracy but also on the speed and efficiency of memory usage. The second suggestion is to develop research in the NLP domain. The survey results that we have obtained are not many that discuss CNN to improve language patterns, recognize name entities, or part of speech recognition. Next, we will test CNN in the case of Name Entity Recognition.

Acknowledgments
Thanks to the Universitas Komputer Indonesia research institute and community service that has funded this research in internal research.

References
[1] Zhang W, Itoh K, Tanida J and Ichioka Y 1990 Parallel distributed processing model with local space-invariant interconnections and its optical architecture Appl. Opt. 29
[2] Heravi H H A E J 2017 Guide to Convolutional Neural Networks : a Practical Application to Traffic-Sign Detection and Classification (Cham, Switzerland: Springer)
[3] Bezak P, Bozek P and Nikitin Y 2014 Advanced Robotic Grasping System Using Deep Learning Procedia Eng. 96 10–20
[4] Krizhevsky A and Hinton G E 2012 ImageNet Classification with Deep Convolutional Neural Networks Proceeding NIPS’12 Proceedings of the 25th International Conference on Neural Information Processing Systems (Lake Tahoe, Nevada) pp 1–9
[5] Liu T, Ma L, Shahroudy A, Chen T, Gu J, Wang X, Wang Z, Cai J, Wang G, Kuen J and Shuai B
2017 Recent advances in convolutional neural networks Pattern Recognit. 77 354–77
[6] Zhang Y and Wallace B 2015 A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification
[7] Widiastuti N I 2018 Deep Learning - Now and Next in Text Mining and Natural Language Processing IOP Conf. Ser. Mater. Sci. Eng. 407
[8] Hu G, Jiang Y, Yan R, Xiao X and Peng S 2017 New deep learning method to detect code injection attacks on hybrid applications J. Syst. Softw. 137 67–77
[9] Karbab E M B, Debbabi M, Derhab A, and Mouheb D 2018 MalDozer: Automatic framework for Android malware detection using deep learning Digit. Investig. 24 S48–59
[10] Martinelli F, Marulli F and Mercaldo F 2017 Evaluating Convolutional Neural Network for Effective Mobile Malware Detection Procedia Comput. Sci. 112 2372–81
[11] Li L, Qin B, Ren W, and Liu T 2017 Document representation and feature combination for deceptive spam review detection Neurocomputing 254 33–41
[12] Ren Y and Ji D 2017 Neural networks for deceptive opinion spam detection: An empirical study Inf. Sci. (Ny). 385–386 213–24
[13] Poria S, Cambria E and Gelbukh A 2016 Aspect extraction for opinion mining with a deep convolutional neural network Knowledge-Based Syst. 108 42–9
[14] Poria S, Peng H, Hussain A, Howard N and Cambria E 2017 Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis Neurocomputing 261 217–30
[15] Liao S, Wang J, Yu R, Sato K and Cheng Z 2017 CNN for situations understanding based on sentiment analysis of twitter data Procedia Comput. Sci. 111 376–81
[16] Karthik V, Nair D and Anuradha J 2018 Opinion Mining on Emojis using Deep Learning Techniques Procedia Comput. Sct. 132 167–73
[17] Severyn A, Moschitti A, Severyn A and Moschitti A 2015 Twitter Sentiment Analysis with Deep Convolutional Neural Networks Proc. 38th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’15 959–62
[18] Gui L, Zhou Y, Xu R, He Y and Lu Q 2017 Learning representations from heterogeneous network for sentiment classification of product reviews Knowledge-Based Syst. 124 34–45
[19] Chen T, Xu R, He Y and Wang X 2017 Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN Expert Syst. Appl. 72 221–30
[20] Zhang S, Grave E, Sklar E and Elhadad N 2017 Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks J. Biomed. Inform. 69 1–9
[21] He B, Guan Y and Dai R 2019 Classifying medical relations in clinical text via convolutional neural networks Artif. Intell. Med. 93 43–9
[22] Carcone A I, Naar-King S, Xu H, Kotov A, Zhu D and Dong M 2016 Text Classification with Topic-based Word Embedding and Convolutional Neural Networks 88–97
[23] Rios A and Kavuluru R 2015 Convolutional neural networks for biomedical text classification 258–67
[24] Sulieman L, Gilmore D, French C, Cronin RM, Jackson G P, Russell M and Fabbri D 2017 Classifying patient portal messages using Convolutional Neural Networks J. Biomed. Inform. 74 59–70
[25] Zheng W, Lin H, Li Z, Liu X, Li Z, Xu B, Zhang Y, Yang Z and Wang J 2018 An effective neural model extracting document level chemical-induced disease relations from biomedical literature J. Biomed. Inform. 83 1–9
[26] Tran T and Kavuluru R 2017 Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks J. Biomed. Inform. 75 S138–48
[27] Li P and Mao K 2019 Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts Expert Syst. Appl. 115 512–23
[28] Wehrmann J, Mattjie A, and Barros R C 2018 Order embeddings and character-level
convolutions for multimodal alignment *Pattern Recognit. Lett.* **102** 15–22

[29] Shi B, Bai X and Yao C 2016 Script identification in the wild via discriminative convolutional neural network *Pattern Recognit.* **52** 448–58

[30] Zhang Y, Er M J and Pratama M 2016 Extractive Document Summarization Based on Convolutional Neural Networks 918–22

[31] Summarization M E, Zhang Y, Member S, Er M J and Member S 2016 Multiview Convolutional Neural Networks for 1–13

[32] Agarwal B, Ramampiaro H, Langseth H and Ruocco M 2018 A deep network model for paraphrase detection in short text messages *Inf. Process. Manag.* **54** 922–37

[33] Bao H, Lu D, Hao H, Hao Y, Xu J, Zheng S and Xu B 2017 Joint entity and relation extraction based on a hybrid neural network *Neurocomputing* **257** 59–66

[34] Xu J, Xu B, Wang P, Zheng S, Tian G, Zhao J, and Xu B 2017 Self-Taught convolutional neural networks for short text clustering *Neural Networks* **88** 22–31

[35] BB L, Wang Q, Wang X and Li W 2017 *Tag Prediction in Social Annotation Systems Based on CNN and BiLSTM* vol 10385 (Springer International Publishing)

[36] Zhang Z, Zou Y and Gan C 2018 Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression *Neurocomputing* **275** 1407–15

[37] Er M J, Zhang Y, Wang N and Pratama M 2016 Attention pooling-based convolutional neural network for sentence modeling *Inf. Sci. (Ny).* **373** 388–403