Indonesian Journal of Electrical Engineering and Computer Science
Vol. 19, No. 2, August 2020, pp. 1000~1009
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v19.i2.pp1000-1009

Indonesian news classification using convolutional neural network

Muhammad Ali Ramdhani[1], Dian Sa'adillah Maylawati [2], Teddy Mantoro[3]
[1]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
[1,2]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[3]Department of Computer Science, Sampoerna University, Indonesia

Article Info

Article history:

Received Dec 20, 2019
Revised Feb 26, 2020
Accepted Mar 11, 2020

Keywords:

Convolutional neural network
Deep learning
Indonesian language process
Natural language processing
Text mining

ABSTRACT

Every language has unique characteristics, structures, and grammar. Thus, different styles will have different processes and result in processed in Natural Language Processing (NLP) research area. In the current NLP research area, Data Mining (DM) or Machine Learning (ML) technique is popular, especially for Deep Learning (DL) method. This research aims to classify text data in the Indonesian language using Convolutional Neural Network (CNN) as one of the DL algorithms. The CNN algorithm used modified following the Indonesian language characteristics. Thereby, in the text pre-processing phase, stopword removal and stemming are particularly suitable for the Indonesian language. The experiment conducted using 472 Indonesian News text data from various sources with four categories: 'hiburan' (entertainment), 'olahraga' (sport), 'tajuk utama' (headline news), and 'teknologi' (technology). Based on the experiment and evaluation using 377 training data and 95 testing data, producing five models with ten epoch for each model, CNN has the best percentage of accuracy around 90,74% and loss value around 29,05% for 300 hidden layers in classifying the Indonesian News data.

Corresponding Author:

Muhammad Ali Ramdhani,
Department of Informatics,
UIN Sunan Gunung Djati Bandung,
Jl. A.H. Nasution 105, Bandung, 40614, Indonesia.
Email: m_ali_ramdhani@uinsgd.ac.id

1. INTRODUCTION

There are almost 7000 languages in the world [1, 2], not including the unidentified local languages, having millions of rules in structure, grammar, even in the form of a letter. This makes an exciting fact to process words and find the insight information or knowledge on it. The technique that specifically processes the language computationally is Natural Language Processing (NLP) [3-5]. NLP can process language with various data, such as speech and text. Usually, the NLP technique combined with Data Mining (DM) or Machine Learning (ML) technique inside. Especially for data text form, NLP can combine with the Text Mining (TM) technology. TM is a technique to discover the insight knowledge from text data, which is unstructured data [6, 7].

In the current TM and NLP research, Deep Learning (DL) method is popular. DL is a development of conventional Artificial Neural Network (ANN) with adding multiple hidden layers between the input layer and output layer [8-10]. Many previous NLP research uses DL method [11, 12], among others: text summarization using DL [13-15], malware classification that uses text data using DL [16], reading text using DL [17], and many others. Many DL algorithms developed for TM or NLP as shown in Figure 1, such as Deep Reinforce Model [18], Deep Unsupervised Learning [13], Convolutional Neural Network (CNN) [17], Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) [12, 19, 20], and other methods.

Most of the NLP research using DL conducted in English. It rarely uses other languages such as Indonesian. Since the Indonesian language has its characteristics, it will be different in data collection, pre-processing, and the result of DL performance in processing the Indonesian language. Therefore, this research utilizes CNN algorithm as one of DL methods to classify the Indonesian News and to evaluate the performance of the accuracy of CNN on the classification result.

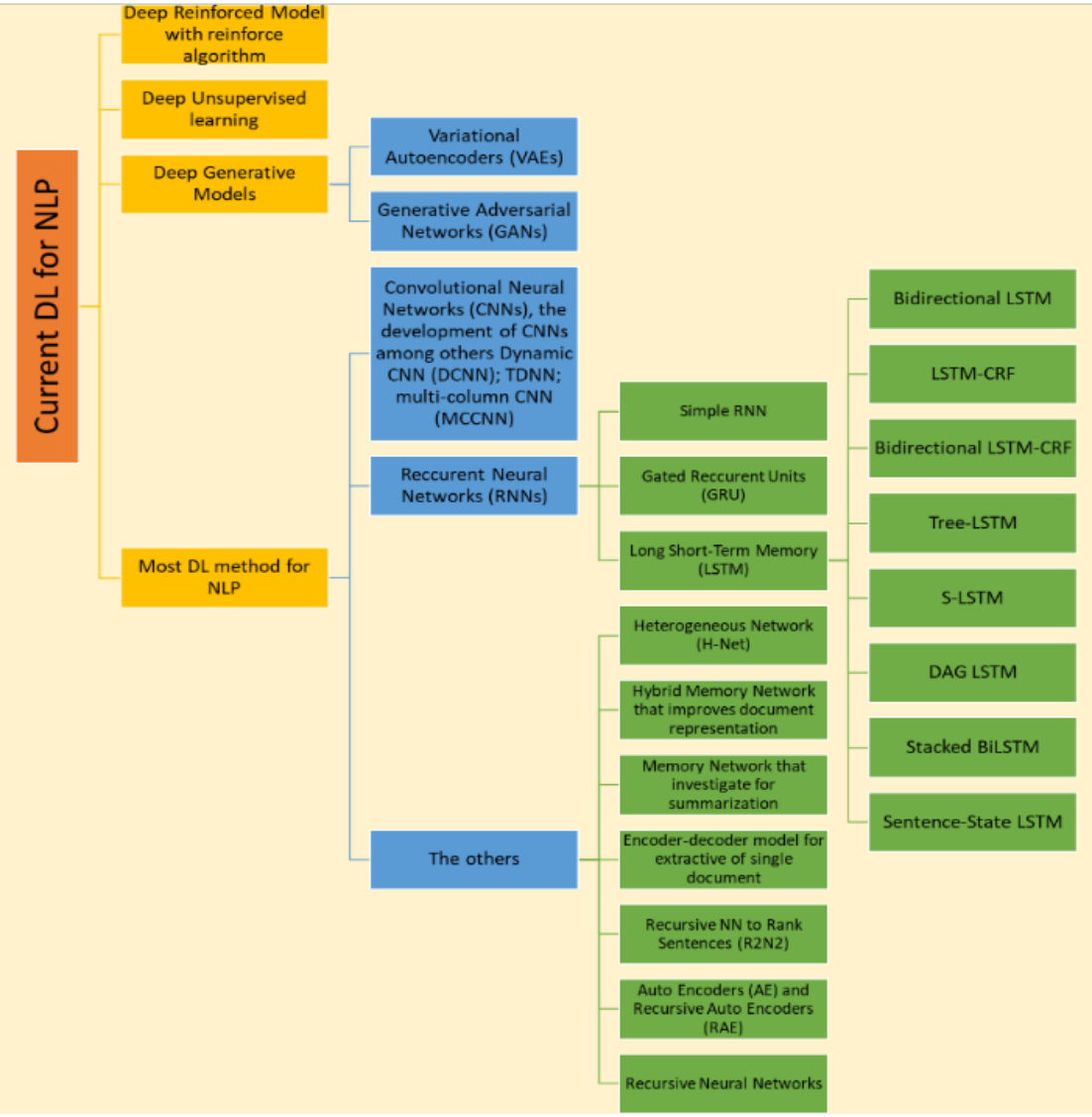

Figure 1. Current DL algorithm being popular for NLP

2. RESEARCH METHOD

Figure 2 describes the flow of activity used in this research as a methodology. Generally, the first process of this research is data collection, where text data collected from various online news websites. Then, Exploratory Data Analysis (EDA) conducted to know the information of data and to make sure that the quality of data is good, clean, and ready for the next process [21, 22]. Text-preprocessing is a critical phase to prepare text data before the mining process, among others: tokenizing, cleaning data, stopword removal, stemming, and changing text from unstructured data into structured text representation [23]. Then, after text-preprocessing, the CNN algorithm will be run to classify the text data with the training and testing process as a classification method in data mining. The last process is the evaluation of the training and testing process, and the accuracy and loss value of each model that is produced. This research used Phython as programming language to conduct the classifiction using CNN.

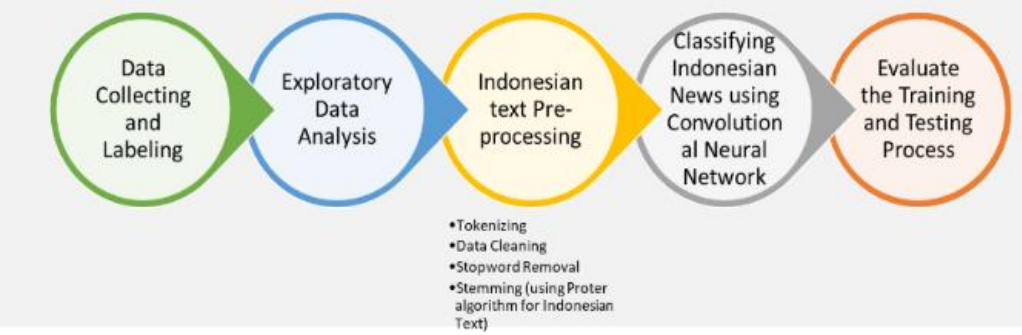

Figure 2. Research activities flow

2.1. Data collection and labeling

Collecting data is conducted manually from several online news websites in the Indonesian Language. The classification process requires labeling for each data so that the Indonesian News data will be labeled based on the type of News content. Those datasets saved in plain text format.

2.2. Exploratory data analysis (EDA)

EDA is a procedure to analyze data so that the data can be analyzed easier, more accurate, more precise with the mathematical method (statistics) processed automatically by machine [24, 25]. EDA provides a summary of numerical data such as mean, median, maximum value, minimum value, and the quartiles. In data analytics, in the case of the DM process, EDA is usually used in the pre-processing process to visualize, to find the missing side, and also to look for a correlation between data or variables. The pre-processing phase is essential for data integration, data selection, data cleaning to improve quality, data transformation, and data reduction to run an efficient mining process.

2.3. Indonesian text preprocessing

Text pre-processing is an important phase to prepare text data well before the mining process. Besides cleaning the text data, text pre-processing can reduce the dimension of data, either through duplication, missing value, or decreasing the number of features [26]. Then, in the text pre-processing, the data text, which is an unstructured data type, will be changed into structured test representation. Many structured text representations include Bag of Words (BoW) [27-31], binary representation [27], n-gram [27, 31], multiple of words (MoW) or multi-words term [27, 32], character and word embedding [12, 33, 34], semantic text representation [35-37], text representation for deep learning [12], symbolic and non-symbolic text representation [38], graph-based text representation [28], active descriptive text representation [34] and document specific representation [39].

2.4. Convolutional neural network (CNN)

CNN is one of the variants of DL algorithms that can take in an input data, assign importance (learnable weights and biases) to various objects in the data, and differentiate one from the other [16, 40]. Same as ANN architecture, CNN consists of neurons that have weight, bias, and activation functions [41, 42]. However, CNN divides the architectures into Feature Extraction Layer and Fully Connected Layer. Feature Extraction Layer is a process for encoding the data to be featured in the form of numbers representing the data that consist of the Convolutional Layer and Pooling Layer. While, Fully Connected Layer has several hidden layers, activation functions, output layers, and loss functions. Feature maps generated from the extraction layer are still in the form of a multidimensional array, so we have to "flatten" or reshape the feature map to be a vector so we can use it as input from Fully Connected Layer. In many research, CNN used for image data.

3. RESULTS AND ANALYSIS

3.1. Indonesian news datasets

Indonesian News datasets collected from various sources such as CNN Indonesia, Merdeka.com, Aneka News, DialySocial.com, and many others. The News datasets labeled into four categories: ‘hiburan’ (entertainment), ‘olahraga’ (sport), ‘tajuk utama’ (headline news), and ‘teknologi’ (technology). All of those data saved in plain text files.

3.2. Result of EDA of Indonesian news dataset

Figure 3 shows the result of EDA process of Indonesian News data collection. Based on the EDA process, from 472 news data, the maximum length of data is 1929, and it can be an outlier data. The range is too far from the mean value of length data, which is 264. From all of the News data collection, there are 12573 vocabularies found.

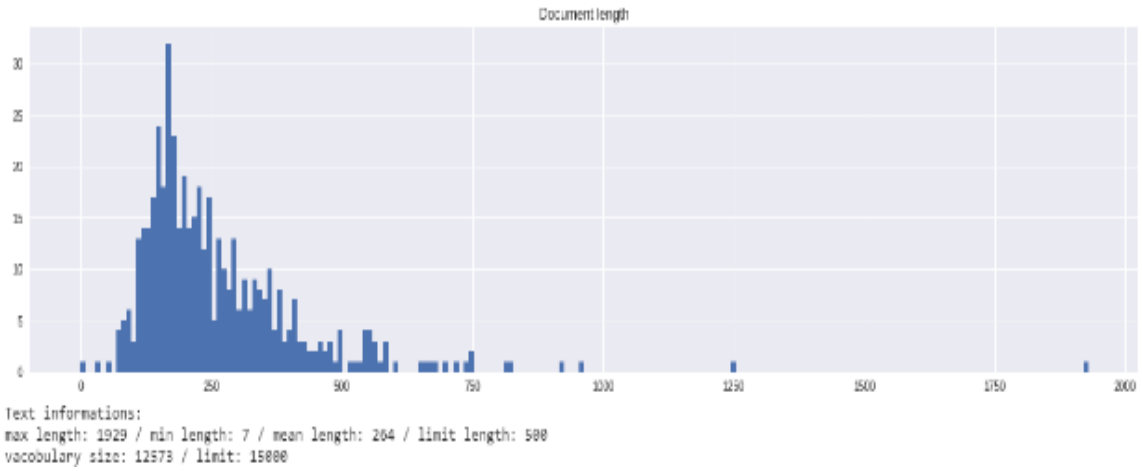

Figure 3. EDA result of indonesian news data collection

3.3. Result of text pre-processing

The pre-processing text data begins with preparing data collection from the real data into clean data text after data cleaning, stopwords removal, and stemming process. For stopwords, it used the modified collection of Indonesian stopwords and Porter Stemming so that the structure and grammar are suitable with the Indonesian language [26, 43]. Table 1 shows the example of the original Indonesian News data that becomes the data after text pre-processing. Following CNN as a DM method that needs embedding representation, this research uses GloVe as embedding representation [44-46]. Figure 4 Figure *4*. The example result of GloVe embedding, training, and testing processshows the slice of the GloVe embedding result as well as the training and testing process of CNN.

Table 1. The example of original data and text pre-processing data

Original Indonesian News Text Data:
Jakarta, CNN Indonesia - PT Honda Prospect Motor (HPM) benar-benar menghentikan produksi mobil multi purpose vehicle (MPV) Freed. Freed berhenti produksi pada Juli 2016 silam, dan pihak Honda tahun ini hanya akan menjual sisa prosuksinya. " Freed sekarang sudah tidak kontinu, bulan Juli terakhir diproduksi, " ujar Direktur Pemasaran dan Layanan Purna Jual PT HPM Jonfis Fandy, belum lama ini. HPM masih menjual Freed, namun dengan angka yang sangat terbatas yang ada di dealer resmi Honda, karena stok dari pabrik sudah benar-benar kosong. Dengan selesainya produksi Freed, membuat Toyota memiliki pasar yang lebih lapang untuk kelas MPV, dan menjadi yang pertama di kelas multi activity vehicle (MAV). Jonfis menegaskan, Honda Indonesia tidak berniat membuka kembali segmen Freed karena konsumennya sudah mulai berkurang. " Segmen itu sudah banyak hilang, diambil oleh HR - V, " ujarnya. Adanya kehadiran Freed terbaru di Jepang tidak membuat mobil ini akan memperpanjang usianya di Indonesia. " Sampai sekarang gak ada rencana. Freed itu 1.000 unit per bulan maksimal. Sementara SUV seperti HR - V bisa 10 ribu per bulan. Rekor tertinggi Freed pas baru diluncurkan 3.000 unit abis itu turun, paling rendah 500 unit, " ujarnya. Dari pengalaman, Honda menilai konsumen Indonesia tidak suka dengan mobil yang memiliki desain terlalu persegi dan kotak. " Konsumen Freed ada tapi tidak banyak. kalau kita kenalkan, terus banting harga, bekasnya orang gak senang. Pada dasarnya orang Indonesia suka mobil yang desainnya stylish, lihat saja per segmen, yang (desain) kotak itu nasibnya jadi pengikut, " ujar Jonsfis. (pit / pit)
Indonesian News Text Data after Text Pre-Processing:
jakarta cnn pt honda prospect hpm henti mult purpose vehicle mpv freed freed hent honda jual prosuks freed akhir produks pasar layan purna pt hpm jonfis fandy hpm jual freed batas dealer honda lesa freed buat toyota pilik mpv jad mult activity vehicle mav jonfis tegas honda niat buka freed konsumen kurang ambil hr ujar hadir freed baru jepang buat panjang usia freed suv hr tingg freed luncur ujar kalam honda tila pilik seg freed nal kas dasar desa stylish nasib ikut jonsfis indonesia motor benar benar produksi mobil produksi juli silam pihak tahun sisa kontinu bulan juli ujar direktur jual angka resmi stok pabrik benar benar kosong produksi pasar lapang kelas pertama kelas indonesia kembali segmen mulai segmen hilang mobil indonesia rencana unit bulan maksimal ribu bulan rekor pas baru unit abis turun rendah unit konsumen indonesia suka mobil desain kotak konsumen terus banting harga orang senang orang indonesia suka mobil lihat segmen desain kotak jadi ujar pit pit

```
Running iteration 1/5
Pretrained embeddings GloVe is loading...
Found 400000 word vectors in GloVe embedding
W0802 04:59:04.520506 140678577579904 deprecation.py:323] From /usr/local/lib/python3.6/dist-packages/tensorflow/python/ops
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where
Train on 377 samples, validate on 95 samples
Epoch 1/10
377/377 [==============================] - 26s 69ms/step - loss: 1.2784 - acc: 0.4589 - val_loss: 1.1958 - val_acc: 0.5053

Epoch 00001: val_loss improved from inf to 1.19578, saving model to model-1.h5
Epoch 2/10
377/377 [==============================] - 24s 64ms/step - loss: 1.2099 - acc: 0.5172 - val_loss: 1.1546 - val_acc: 0.5053

Epoch 00002: val_loss improved from 1.19578 to 1.15458, saving model to model-1.h5
Epoch 3/10
377/377 [==============================] - 24s 64ms/step - loss: 1.1626 - acc: 0.5172 - val_loss: 1.1170 - val_acc: 0.5053

Epoch 00003: val_loss improved from 1.15458 to 1.11703, saving model to model-1.h5
Epoch 4/10
377/377 [==============================] - 24s 64ms/step - loss: 1.1016 - acc: 0.5172 - val_loss: 1.0561 - val_acc: 0.5053

Epoch 00004: val_loss improved from 1.11703 to 1.05609, saving model to model-1.h5
Epoch 5/10
377/377 [==============================] - 24s 64ms/step - loss: 1.0275 - acc: 0.5199 - val_loss: 0.9520 - val_acc: 0.5684
```

Figure 4. The example result of GloVe embedding, training, and testing process

3.4. Training and testing result of CNN

Training and testing process of CNN algorithms to classify Indonesian News are conducted with several scenarios, as follows:

1. The experiment uses Phyton programming language with TensorFlow and Keras package used for NLP and DL.
2. From 472 News text data in total, 377 used for the training process, and 95 used for the testing process.
3. CNN algorithm produces five models with each model having ten epochs.
4. The total hidden layers for the experiment are 100, 200, and 300.
5. The activation function used in this experiment is Sigmoid.
6. Embedding representation for text data used in this experiment is GloVe.

Figure 5-10 and Table 2 provide graphics of evaluation, training, and validation process. Then, Table 3 and Figure 11 show the evaluation result for the testing process. The analysis of the experiment explained in section 3.5 in more detail.

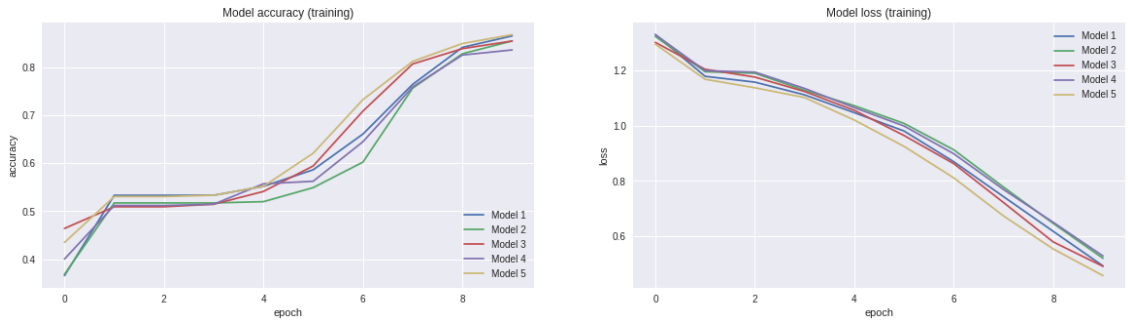

Figure 5. Graphics result of training process with 100 hidden layers

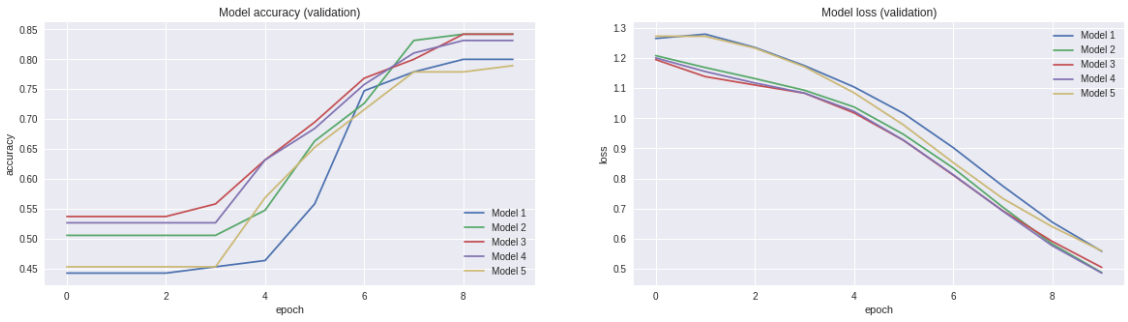

Figure 6. Graphics result of validation of training process with 100 hidden layers

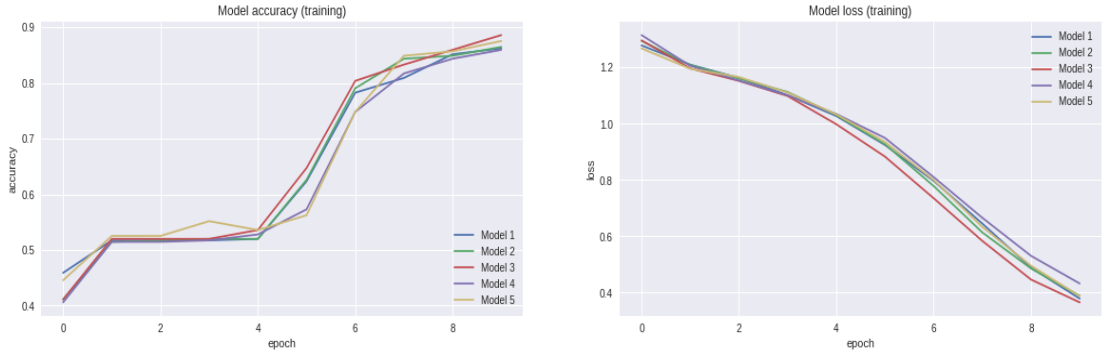

Figure 7. Graphics result of the training process with 200 hidden layers

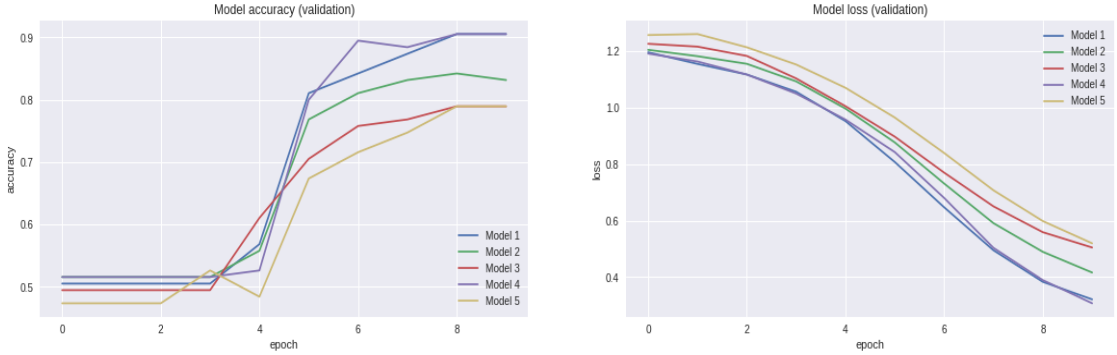

Figure 8. Graphics result of validation of training process with 200 hidden layers

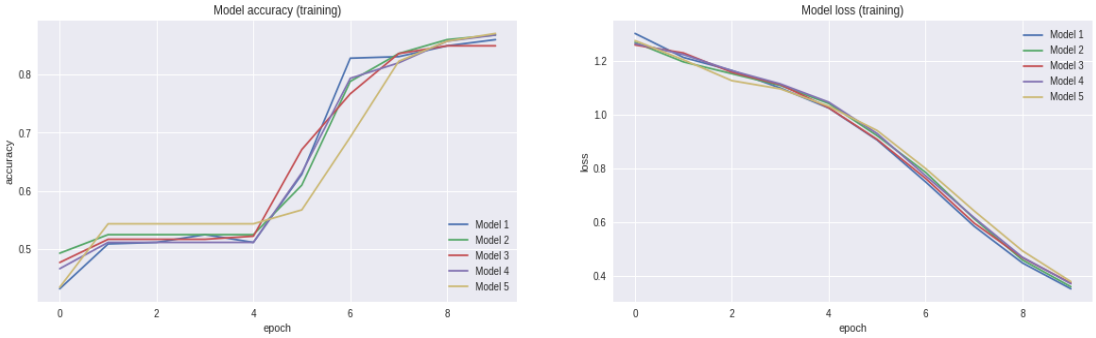

Figure 9. Graphics result of training process with 300 hidden layers

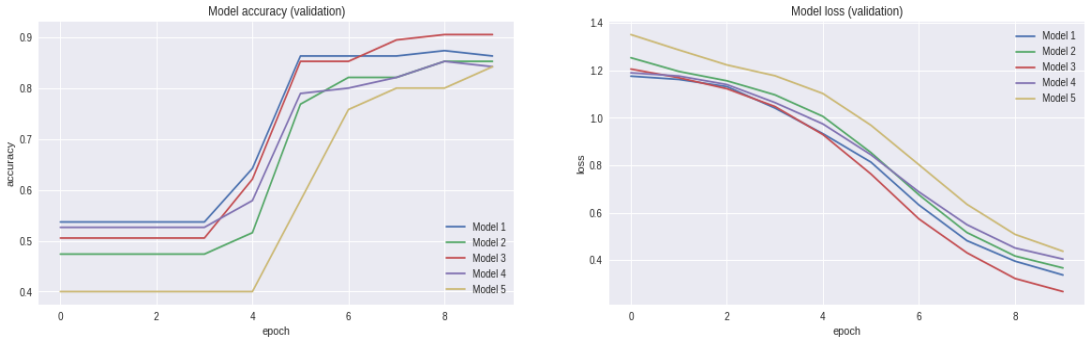

Figure 10. Graphics result of validation of training process with 300 hidden layers

Table 2. Training and validation evaluation result of accuracy and loss value of CNN to classify Indonesian news text data

Hidden Layer	Training Process		Validation Process	
	Accuracy	Loss Value	Accuracy	Loss Value
100	0,8136	0,3586	0,8	0,4126
200	0,8136	0,3674	0,8	0,4372
300	0,7895	0,486	0,7895	0,5194

Table 3. Testing evaluation result of accuracy and loss value of CNN to classify Indonesian news

Model	100 Hidden Layer		200 Hidden Layer		300 Hidden Layer	
	Accuracy	Loss Value	Accuracy	Loss Value	Accuracy	Loss Value
Model 1	0,8136	0,3586	0,8	0,4126	0,9158	0,2766
Model 2	0,8136	0,3674	0,8	0,4372	0,9053	0,2813
Model 3	0,8136	0,3564	0,8	0,4463	0,9053	0,2785
Model 4	0,7895	0,399	0,8	0,4399	0,9053	0,2755
Model 5	0,7895	0,486	0,7895	0,5194	0,9053	0,3406
Average	0,80396	0,39348	0,7979	0,45108	0,9074	0,2905

3.5. Analysis and evaluation result

Based on the result of the experiment, this research finds that:

a) CNN algorithm is not only used for image data but also can used for text data. In this case, for the Indonesian text data. The result of the experiment proves that CNN can classify the Indonesian News well with the average of testing result around 90,74%. It is different from clusterization that does not focus on accuracy but has many interpretations [47-50]. On the other hand, CNN, as the classification, is too complex to analyze and interpret the result, including to see the relationship between variables or parameters and to identify which set of parameters gives the highest impact. However, in harmony with much previous research, DL still has the highest accuracy, and it is efficient for DM, ML, and also NLP [11, 12].

b) Table 2 shows that more hidden layers tend to decrease the accuracy of the training process and validation process even though it is not significant. In contrast, the loss value of the training process increases. This result is in comparison with the test results shown in Table 3, where the largest hidden layer has the highest accuracy. Based on these results, it can state that the higher hidden layers will be better for accuracy. However, this result also can be influenced by the quality of data, EDA (for example many outlier data) and text pre-processing phase (such as an uncomplete dictionary), and another parameter setting.

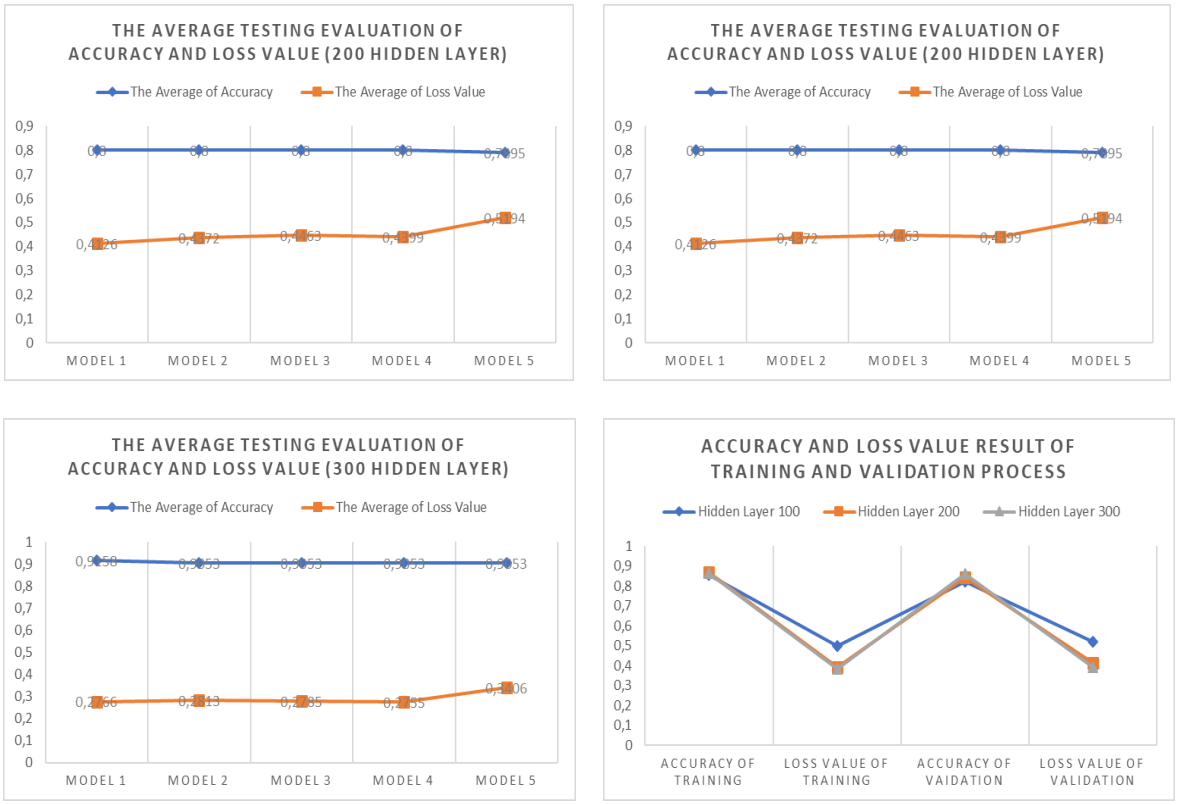

Figure 11. Graphics result of the testing process and the average evaluation of the training and validation process

4. CONCLUSION

NPL research can produce influencing results by type of language. For example, different treatment of English and the Indonesian language will give different results. The NPL can also combine with Data Mining, Machine Learning, and Text Mining used for text data. In the current NLP research, DL as one of the Data Mining methods is popular. This study is successful in using CNN, which is one of DL algorithms, in classifying the Indonesian News text data, where usually CNN is used for image data. The result of the experiment shows that CNN has considered high accuracy in classifying the Indonesian News.

For further research, it can use more data collection that represents big data. Deep Learning method appears because of the needs of the big data era. Then, the experiment should be enhanced, such as by the development of the variation of hidden layers, production of various models, and difference of dataset division between training data and testing data. Besides, the next research can use another activation function besides Sigmoid, another embedding representation besides GloVe, even another Deep Learning algorithm besides CNN.

REFERENCES

[1] Alyosha, "Number of languages in the world (in Bahasa)," 2016. [Online]. Available: https://alyoshaminded.wordpress.com/tag/jumlah-bahasa-di-dunia/

[2] Republika, "Mapped! number of languages worldwide, where is Indonesia language? (in Bahasa) 2016. [Online]. Available: https://www.republika.co.id/berita/internasional/global/15/12/29/o02mbk366-terpetakan-jumlah-bahasa-di-seluruh-dunia-dimana-posisi-indonesia

[3] P. M. Nadkarni, *et al*, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544-551, 2011.

[4] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51-89, 2003.

[5] Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261-266, 2015.

[6] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Elsevier, 2012.

[7] Y. E. Zohar, "Introduction to text mining," *Automated Learning Group, University of Illinois*, 2002.

[8] L. Deng, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197-387, 2014.

[9] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85-117, 2015.

[10] J. Ahmad, H. Farman, and Z. Jan, "Deep learning methods and applications," *Deep Learning: Convergence to Big Data Analytics*, pp. 31-42, 2019.

[11] A. Kulkarni and A. Shivananda, "Deep learning for NLP," in *Natural Language Processing Recipes*, pp. 185-227, 2019.

[12] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, 2018.

[13] M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," *Expert Syst. Appl.*, vol. 68, pp. 93-105, 2017.

[14] G. Rossiello, "Neural abstractive text summarization," in *CEUR Workshop Proceedings*, vol. 1769, pp. 70-752016.

[15] M. Patel, A. Chokshi, S. Vyas, and K. Maurya, "Machine learning approach for automatic text summarization using neural networks," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 7, no. 1, pp. 194–202, 2018.

[16] M. Kalash, M. Rochan, N. Mohammed, N. D. B. Bruce, Y. Wang, and F. Iqbal, "Malware Classification with Deep Convolutional Neural Networks," *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Paris, pp. 1-5, 2018.

[17] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[18] K. Yao, L. Zhang, T. Luo, and Y. Wu, "Deep reinforcement learning for extractive document summarization," *Neurocomputing*, vol. 284, pp. 52-56, 2018.

[19] M. Day and C. Chen, "Artificial intelligence for automatic text summarization," *2018 IEEE Int. Conf. Inf. Reuse Integr.*, pp. 478–484, 2018.

[20] Y. Zhang, Q. Liu, and L. Song, "Sentence-state lstm for text representation," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 317-327, 2018.

[21] J. T. Behrens, "Principles and procedures of exploratory data analysis," *Psychol. Methods*, vol. 2, no. 2, pp. 131, 1997.

[22] M. Stuart, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, "Understanding robust and exploratory data analysis.," *Stat.*, 2006.

[23] D. S. Maylawati and G. A. P. Saptawati, "Set of frequent word item sets as feature representation for text with Indonesian slang," *Journal of Physics Conference Series*. vol. 801. no. 1. 2017.

[24] J. W. Tukey, "The future of data analysis," *The annals of mathematical statistics*, vol. 33, no. 1, pp. 1-67, 1962.

[25] D. C. Hoaglin, "John W. Tukey and data analysis," *Quality control and applied statistics*, vol. 49, no. 5, pp. 549-552, 2004.

[26] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, "Flexibility of Indonesian text pre-processing library," *Indonesian Journal of Electrical Engineering and Computer Science*., vol. 13, no. 1, pp. 420–426, 2019.
[27] B. S. Harish, D. S. Guru, and S. Manjunath, "Representation and classification of text documents: A brief review," *IJCA, Spec. Issue Recent Trends Image Process. Pattern Recognit*., no. 2, pp. 110–119, 2010.
[28] W. Jin and R. K. Srihari, "Graph-based text representation and knowledge discovery," in *Proceedings of the 2007 ACM symposium on Applied computing*, pp. 807-811, 2007.
[29] W. Pu, N. Liu, S. Yan, J. Yan, K. Xie, and Z. Chen, "Local word bag model for text categorization," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, pp. 625-630, 2007.
[30] H. M. Wallach, "Topic modeling: beyond bag-of-words," *Proceedings of the 23rd international conference on Machine learning*, pp. 977–984, 2006.
[31] A. Sethy and B. Ramabhadran, "Bag-of-word normalized n-gram models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1594–1597, 2008.
[32] D. S. Maylawati, M. A. Ramdhani, A. Rahman, and W. Darmalaksana, "Incremental technique with set of frequent word item sets for mining large Indonesian text data," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–6, 2017.
[33] M. Kageback, O. Mogren, N. Tahmasebi, and D. Dubhashi, "Extractive summarization using continuous vector space models," *Proc. 2nd Work. Contin. Vector Sp. Model. their Compos.*, pp. 31-39, 2014.
[34] Y. Zhang, M. Lease, and B. C. Wallace, "Active discriminative text representation learning," *Proc. Thirty-First AAAI Conf. Artif. Intell.*, pp. 3386–3392, 2016.
[35] K. Zupanc and Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Syst.*, vol. 120, pp. 118-132, 2017.
[36] S. T. Dumais, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2005.
[37] J. Y. Yeh, H. R. Ke, W. P. Yang, and I. H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Inf. Process. Manag.*, vol. 41, no. 1, pp. 75–95, 2005.
[38] H. Schuetze, H. Adel, and E. Asgari, "Nonsymbolic text representation," *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguist. Comput. Linguis*t., vol. 1, pp. 785–796, 2017.
[39] K. N. Singh and H. M. Devi, "Document representation techniques and their effect on the document clustering and classification: a review," *Int. J. Adv. Res. Comput. Sci*., vol. 8, no. 5, pp. 1780–1784, 2017.
[40] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 2414-2423, 2016.
[41] S. Sena, "Introduction deep learning part 7: convolutional neural network (CNN)(in Bahasa)," 2017, Online]. Available: https://medium.com/@samuelsena/pengenalan-deep-learning-part-7-convolutional-neural-network-cnn-b003b477dc94.
[42] I. W. Suartika E. P, "Image classification using convolutional neural network (CNN) at Caltech 101(in Bahasa)," *J. Tek. IT*S, vol. 5, no. 1, pp. 76, 2016.
[43] D. S. Maylawati, W. B. Zulfikar, C. Slamet, and M. A. Ramdhani, "An improved of stemming algorithm for mining Indonesian text with slang on social media," in *6th International Conference on Cyber and IT Service Management (CITSM 2018)*, pp. 1-6, 2018.
[44] H. Zamani and W. B. Croft, "Relevance-based word embedding," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 505-514, 2017.
[45] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative study of word embedding methods in topic segmentation," *Procedia Computer Science*, vol. 112, pp. 340-349, 2017.
[46] M. Kamkarhaghighi and M. Makrehchi, "Content tree word embedding for document representation," *Expert Syst. Appl.*, vol. 90, pp. 241-249, 2017.
[47] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97-107, 2014.
[48] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edi. United States of America: Morgan Kaufman, 2012.
[49] F. A. Hermawati, *Data Mining*, Yogyakarta: Andi, 2013.
[50] A. Lamani, B. Erraha, M. Elkyal, and A. Sair, "Data mining techniques application for prediction in OLAP cube," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 3, pp. 2094-2102, 2019.

BIOGRAPHIES OF AUTHORS

Muhammad Ali Ramdhani is a Professor in Information Technology Research in the Department of Informatics at the UIN Sunan Gunung Djati Bandung, Indonesia. His current research interests focus on Information System, Expert System, Decision Support System, Strategic Management, and Research Methodology.

Dian Sa'adillah Maylawati is a lecturer in the Department of Informatics at the UIN Sunan Gunung Djati Bandung, Indonesia. Her current research interests focus on Software Engineering, Expert System, Text Mining, and Natural Language Processing. She takes the Ph.D. degree in Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM).

Teddy Mantoro is a Computer Science Professor at Sampoerna University, Jakarta. He obtained a Ph.D., an MSc and a BSc, all in Computer Science and his Ph.D. awarded from the School of Computer Science, the Australian National University (ANU), Canberra, Australia. He is a Senior Member of IEEE. His research interest is in information Security, pervasive/ ubiquitous computing, wireless sensor network, context-aware computing, mobile computing, and intelligent environment/ IoT.