:: IJEEIT ::
(International Journal of Electrical Engineering and Information Technology)
Volume 04 Number 02 September 2021

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

Implementation of Deep Learning to Detect Indonesian Hoax News with Convolutional Neural Network Method

Cheevin Yoviananda
Faculty of Computer Science, Narotama University, Indonesia
edu4cheexs@gmail.com

Tresna Maulana Fahrudin
Faculty of Computer Science, University of Pembangunan Nasional “Veteran” East Java, Indonesia
tresna.maulana.ds@upnjatim.ac.id

ABSTRACT

This study aims to establish and test a model that is used to determine valid news and hoax news. The method used is the Convolutional Neural Network (CNN) method and Word2Vec as embeddings. The research stages consist of data collection, pre-processing, word embeddings, model formation, and testing the results obtained. The data used is 958 news. After testing with the distribution of data by 80% as training data and 20% as test data and 5 times epoch, the model that has been formed can determine valid news and hoax news well. In this study, a model with a vector dimension of 400 as input data and a multiple filter size of 3,4,5 became the best model. The resulting accuracy, precision, and recall are 0.91. These results are influenced by the selection of the size of the vector dimensions on the output of Word2Vec, the selection of the filter size on the convolution layer, and the addition of the Indonesian Wikipedia corpus into the corpus used.

Keywords:
Convolutional Neural Network, Corpus, Embeddings, Word2Vec

1. Introduction

News or information that does not match the facts or can also be referred to as hoax news has been spread through websites, news portals, and social media. In 2017, reports on hoax news were bigger than reports on pornography cases. The number of reports in January 2017 regarding hoax news was 5,070 complaints, while regarding pornography, there were 308 complaints (Damar, 2017). According to Henri Septanto (Septanto, 2018), hoax news develops due to various factors, such as political interests, the public has not been able to filter the information obtained and it has become a business field for personal interests. This can be said to be very worrying because news from traditional media and social media has an important role in the socio-politics of each individual (Oshikawa et al., 2018). One example of hoax news is news about a tsunami that will occur in the coastal city of Kupang so that residents flock to higher places to take shelter (Apriyono, 2021). The residents of Polewali Mandar, West Sulawesi have experienced the same thing (CNNIndonesia, 2018). Another example is the news about canned milk which is believed to speed up the recovery of someone with COVID-19 symptoms. Residents buy in a frenzy, causing a shortage of goods (Azizah, 2021). Various efforts have been made to prevent the spread of hoax news, such as direct fact-checking by the MAFINDO (Masyarakat Anti Fitnah Indonesia) community to utilize artificial intelligence by looking for patterns in hoax news. Several studies have been carried out to prevent the spread of hoax news, such as that conducted by Aulia and Julio (Afriza and Adisantoso, 2018). They analyze hoax news obtained online (websites, social media) using the Rocchio classification approach and Naive Bayes multinominal. The data used consists of 300 hoax news stories and 300 valid news stories. The results of the study stated that the Rocchio approach had better accuracy than the Naive Bayes multinominal approach in classifying hoax news, with a difference in the accuracy of 17.666%. The weakness of the research is in the preprocessing. Each word has to go through a standardization stage, so it takes time if the data used is too much, and it performs a search if new slang words are found. Antonius and Metty (Kurniawan and Mustikasari, 2021) conducted a study to determine hoax news using the Convolutional Neural Network (CNN) and

Long Short Term Memory (LTSM) methods. The data used consists of 802 valid news stories and 984 hoax news stories. The data was obtained from a website that has provided hoax news and valid news. The Word Embedding process uses Word2Vec with the corpus provided by the Wiki. The results of the study stated that the CNN and LTSM methods can determine hoax news and valid news well, with an accuracy of 0.88 and 0.84, respectively. This research does not explain the kernel and input vector dimensions used in the convolution layer. Based on the background, description of the problem, and previous research, the purpose of this study is to determine hoax news and valid news using the CNN method by finding the kernel size and input vector dimensions to get maximum results.

2. Research Method

The system design in this study consists of several flows that represent the input, process, and output of the system. The system design in this study is shown in Figure 1.

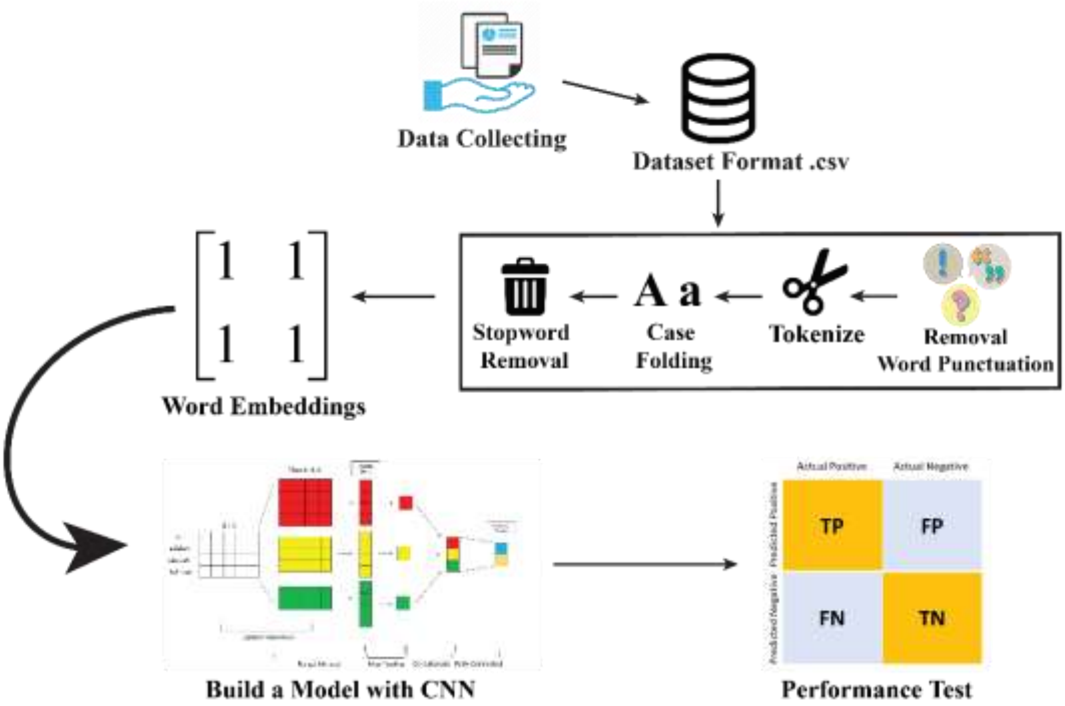


Figure 1. Design System Plan

2.1 Data Collection

The data used in this study was taken manually by researchers from January 29, 2020, to May 31, 2020, regarding news in the Indonesian language related to the topic of the coronavirus, both hoaxes and facts from several websites, namely turnbackhoax, fact check, and tempo check. The data collected consisted of 958 samples and was divided into 534 factual samples and 424 hoax samples. The data will be saved into a file with the extension CSV. The dataset used is news data and news labels. The news label consists of values 0 and 1. A value of 0 is for presenting hoax news and a value of 1 is for factual news.

2.2 Pre-processing

The data that has been collected will be processed in the preprocessing stage. This stage aims to obtain data with a good representation so that it can meet the right data (Hidayatillah et al., 2019). The following steps are carried out during preprocessing.

2.2.1 Punctuation Removal

The files that have been collected will go through the first stage, namely the stage of deleting punctuation marks (!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~) so that the data will be used as input data only in the form of an alphabet.

2.2.2 Tokenizing

After going through the punctuation removal stage, the sentences in each document will be broken down into words which are commonly called tokens.

2.2.3 Case Folding

At this stage, the entire document is converted to lowercase. This stage needs to be done to avoid double sentences due to differences in the way they are written.

2.2.4 Stopword Removal

The last stage in preprocessing is stopword removal. Stopwords are common words that usually appear but have no meaning. At this stage, the words included in the stopword list will be deleted, such as the words "di", "ke", "dan", "ini", and "karena". In this study, researchers used a python library called Sastrawi to help carry out the stopword removal process.

Table 1. Preprocessing stage

Stage	Input	Output
Punctuation Removal	Di Indonesia, selain 60% dari rokok yang kita hisap adalah pajak untuk negara, ternyata perokok tidak disukai Covid-19.	Di Indonesia selain 60 dari rokok yang kita hisap adalah pajak untuk negara ternyata perokok tidak disukai Covid19
Tokenizing	Di Indonesia selain 60 dari rokok yang kita hisap adalah pajak untuk negara ternyata perokok tidak disukai Covid19	[‘Di’, 'Indonesia', ‘selain’, '60', ‘dari’ 'rokok', ‘yang’, ‘kita’, 'hisap', ‘adalah’, 'pajak', ‘untuk’, 'negara', 'ternyata', 'perokok', ‘tidak’, 'disukai', 'Covid19']
Case Folding	[‘Di’, 'Indonesia', ‘selain’, '60', ‘dari’ 'rokok', ‘yang’, ‘kita’, 'hisap', ‘adalah’, 'pajak', ‘untuk’, 'negara', 'ternyata', 'perokok', ‘tidak’, 'disukai', 'Covid19']	[‘di’, 'indonesia', ‘selain’, '60', ‘dari’ 'rokok', ‘yang’, ‘kita’, 'hisap', ‘adalah’, 'pajak', ‘untuk’, 'negara', 'ternyata', 'perokok', ‘tidak’, 'disukai', 'covid19']
Stopword Removal	[‘di’, 'indonesia', ‘selain’, '60', ‘dari’ 'rokok', ‘yang’, ‘kita’, 'hisap', ‘adalah’, 'pajak', ‘untuk’, 'negara', 'ternyata', 'perokok', ‘tidak’, 'disukai', 'covid19']	['indonesia', '60', 'rokok', 'hisap', 'pajak', 'negara', 'ternyata', 'perokok', 'disukai', 'covid19']

2.2.5 Word Embeddings

Word Embedding aims to convert a word into a vector-based on its occurrence with other words. Embeddings generally represent the geometric coding of words based on how often they appear together in the text corpus (X. Zhang et al., 2015). The word embedding stage in this study uses the Word2Vec approach with the Skip-gram model (Mikolov et al., 2013). The corpus used comes from a collection of news stories that have been collected and Wiki Indonesia. The vector dimensions generated at this stage are 100, 200, 300, 400, 500.

2.3 Convolutional Neural Network

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

The Convolutional Neural Network (CNN) is one of the developments of the Multilayer Perceptron (MLP) which is used to solve certain problems. The CNN model design in this study can be seen in Figure 2.

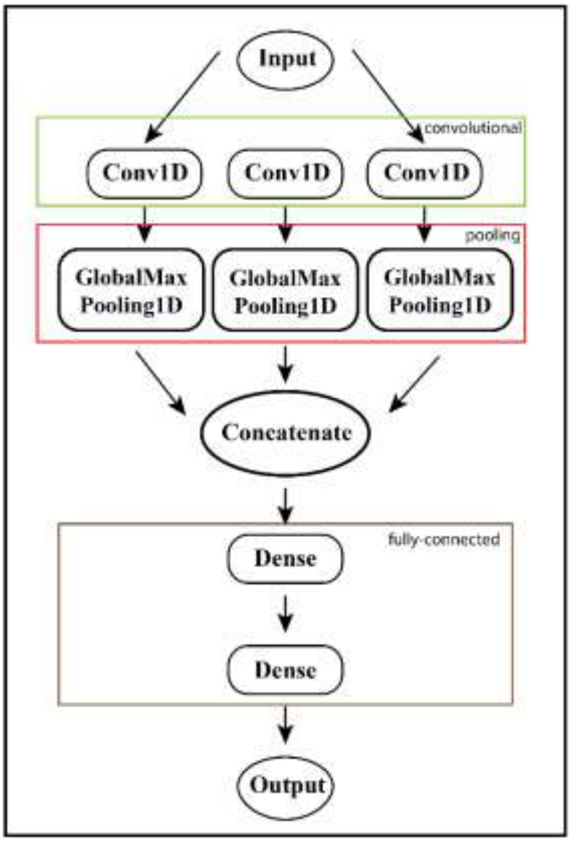


Figure 2. CNN Model Architecture

2.3.1 Data Input

The CNN model's input data is in the form of a document, each of whose words has been processed using Word2Vec. The number of sentence lengths varies from text to document. CNN requires input data with the same amount of sentence lengths in each document, hence papers with insufficient sentence lengths must be supplemented with zero vectors. The number of sentence lengths in each will be changed for this study based on the maximum length of a document. A maximum of 237 words can be found in the dataset that was gathered. The data will be split into two portions, training data, and test data, with an 80:20 ratio.

2.3.2 Convolutional Layer

The convolution layer consists of a set of filters or kernels (Zhong et al., 2019). In this layer, the input data vector (W) whose sentence length (h) has been equalized will be multiplied by the filter (ω) of the specified size. This can be seen in equation 1.

$$c_i = f(\omega . W_{i:i+h-1} + b) \quad (1)$$

The symbol f is an activation function and b is a bias. In this study, the kernel size used in the convolution layer refers to the advice given by Ye Zhang and Byron in their research (Y. Zhang and Wallace, 2015) by finding the best kernel for the data that has been collected and the output dimensions of 128. The output of this layer can be called a feature. folder.

2.3.3 Activation Function

The function used in the convolution layer is ReLu (Rectified Linear Unit). This refers to the results of research by Zhang and Byron (Y. Zhang and Wallace, 2015). ReLu serves as a filter for each neuron. If there is a neuron that has a negative value, then ReLu will change that value to 0. The graphical form of the ReLu function can be seen in Figure 3.

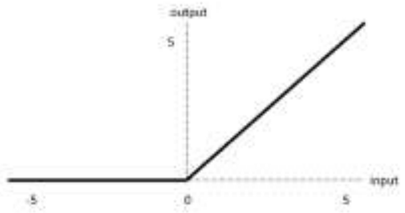
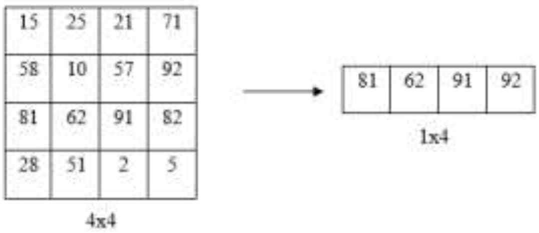


Figure 3. ReLu Function

2.3.4 Subsampling Layer

Subsampling aims to reduce the size of the vector dimensions of the feature map data. The Global Max Pooling method will be used in this study. The way Global Max Pooling works is almost the same as Max pooling. The difference between the two methods is in the output. The output of Global Max Pooling has the same size as the feature map size. An example of how the Global Max Pooling method works can be seen in Figure 4.



15	25	21	71
58	10	57	92
81	62	91	82
28	51	2	5

4x4

81	62	91	92

1x4

Figure 4. An Example of Global Max Pooling

2.3.4 Concatenate Layer

In this layer, the output from the subsampling layer will be combined into 1. Once combined, the dimension size will be changed to 1 dimension so that it can be processed in the next layer.

2.3.5 Fully-connected Layer

The fully-connected layer will go through the backpropagation process to get the best weight so that it can determine the label that has been determined. In this research CNN model design, there are 2 dense pieces. The first Dense function is to receive input from the concatenate layer and store values that have gone through weight calculations. The second dense layer serves to classify news based on the value of the first dense. The researcher also uses the dropout method, the Adam Optimization algorithm, and the Binary Cross Entropy Loss Function in this layer. The application of the dropout method and Adam's algorithm was inspired by research conducted by Xiaoyilei et al (Yang et al., 2019) to obtain optimal accuracy in the classification process.

2.4 Performance Test

At this stage, the model that has been created will be tested and evaluated to determine its performance. The Confusion Matrix will be used for model evaluation. The number of successful and failed data predicted by the model will be stored in one Cofusion Matrix table (Pratama et al., 2020). The form of the Confusion Matrix can be seen in Table 2.

Table 2. Confusion matrix

	Aktual Positif	Aktual Negatif
Prediksi Positif	TP	FP
Prediksi negatf	FN	TN

By utilizing the confusion matrix, we can calculate accuracy, precision, and recall. This is to determine the performance of the model that has been made. The equations of accuracy, precision, and recall can be seen in equations II, III, IV.

$$\text{Accurate} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{(II)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(III)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(IV)}$$

3. Result and discussion

3.1 The Result of the Vector Dimension Size Experiment on the Input

The output of Word2Vec is a vector model with dimensions of 100,200,300,400,500. Some of these models will be tested by CNN with kernel 3,4,5. This experiment and others have an epoch of 5. Table 3 shows the results of testing the CNN model with different dimensions of the input data vector. The CNN model with vector dimensions of 400 has the highest accuracy of 0.91 and the model with the lowest accuracy is found in the model with vector dimensions of 100 with an accuracy of 0.86. Based on the test results in Table 3, this study will use input data with a vector dimension of 400.

Table 3. The result of the vector dimension size experiment

Size Vector Dimension	*Precision*	*Recall*	Accuracy
100	0.89	0.87	0.86
200	0.84	0.93	0.88
300	0.89	0.87	0.87
400	0.91	0.91	0.91
500	0.87	0.91	0.88

3.2 Results of the Filter Size Experiment on the convolution layer

Ye Zhang and Byron suggested trying the single filter first. The recommended single filter size is between 1 and 10. In Table 4, the model with a single filter of 3 has the highest accuracy of 0.90 and the model with a single filter of 5 has the lowest accuracy with an accuracy of 0.86. After finding a single filter with the best accuracy, we tried further to combine single filters into multiple filters.

Table 4. Model performance results based on single filter

Filter Size	Precision	Recall	Accuracy
1	0.89	0.88	0.88
3	0.92	0.89	0.90
5	0.96	0.76	0.86
7	0.86	0.95	0.89

9	0.88	0.92	0.89

Table 5. Model performance results based on single filter combination

Filter Size	Precision	Recall	Accuracy
2, 3, 4	0.86	0.94	0.89
3, 4, 5	0.91	0.91	0.91
4, 5, 6	0.83	0.95	0.87
5, 6, 7	0.83	0.94	0.86
7, 8, 9	0.78	0.99	0.78

The model with a filter size of 3,4,5 has the highest accuracy of 0.91 and the model with a filter size of 7,8,9 has the lowest accuracy of 0.78. This is in accordance with what Ye Zhang and Byron explained about the combination or multiple filters that can function optimally based on the best single filter around. So the best model for the data obtained is a model with a multiple filter size of 3,4,5 with Word2Vec input data of 400.

3.3 The Effect of Adding Indonesian Wikipedia

It has been mentioned earlier that we have added the Indonesian Wikipedia corpus to the corpus we are using. It has been commonly used in similar studies. However, we just want to know how much impact the addition of the Indonesian Wikipedia corpus will have.

Table 6. Model performance results based on corpus addition

	Before	After
Recall	0.87	0.91
Precision	0.74	0.91
Accuracy	0.77	0.91

The addition of the Indonesian Wikipedia corpus has a significant effect on maximizing the model's performance. There is a fairly large difference between the accuracy produced by the model without the addition of a corpus and the model that has gone through the addition of a corpus, which is 0.14.

4. Conclusion

Based on this study, the use of the CNN method was successfully applied. The model can determine hoax news and valid news very well. The model gets an accuracy, recall, and precision score of 0.91. This is influenced by the selection of dimensions for the input data, a good filter on the convolution layer, and the addition of a corpus in the formation of Word2Vec to maximize the performance of the model in determining news.

REFERENCES

Afriza, A., & Adisantoso, J. (2018). Metode Klasifikasi Rocchio untuk Analisis Hoax. *Jurnal Ilmu*

Komputer Dan Agri-Informatika, *5*(1), 1. https://doi.org/10.29244/jika.5.1.1-10

Apriyono, A. (2021). *Warga di Pesisir Kota Kupang Panik Termakan Hoaks Bakal Ada Tsunami*. Liputan6. https://www.liputan6.com/regional/read/4525533/warga-di-pesisir-kota-kupang-panik-termakan-hoaks-bakal-ada-tsunami

Azizah, K. N. (2021). *Susu Beruang Habis Diborong Warga +62, Ini Kata Nestle*. DetikHealth. https://health.detik.com/berita-detikhealth/d-5631596/susu-beruang-habis-diborong-warga-62-ini-kata-nestle

CNNIndonesia. (2018). *Warga Panik Akibat Hoaks Gempa-Tsunami di Sulbar*. https://www.cnnindonesia.com/nasional/20181001165424-24-334695/video-warga-panik-akibat-hoaks-gempa-tsunami-di-sulbar

Damar, A. M. (2017). *Jumlah Aduan Hoax dan SARA Lampaui Konten Pornografi*. Liputan6. https://www.liputan6.com/tekno/read/3053599/jumlah-aduan-hoax-dan-sara-lampaui-konten-pornografi

Hidayatillah, R., Mirwan, M., Hakam, M., & Nugroho, A. (2019). Levels of Political Participation Based on Naive Bayes Classifier. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(1), 73. https://doi.org/10.22146/ijccs.42531

Kurniawan, A. A., & Mustikasari, M. (2021). Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia. *Jurnal Informatika Universitas Pamulang*, *5*(4), 544. https://doi.org/10.32493/informatika.v5i4.6760

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). *A Survey on Natural Language Processing for Fake News Detection*. http://arxiv.org/abs/1811.00770

Pratama, A. R., Mustajib, M., & Nugroho, A. (2020). Deteksi Citra Uang Kertas dengan Fitur RGB Menggunakan K-Nearest Neighbor. *Jurnal Eksplora Informatika*, *9*(2), 163–172. https://doi.org/10.30864/eksplora.v9i2.336

Septanto, H. (2018). Pengaruh Hoax dan Ujaran Kebencian Sebuah Cyber Crime dengan Teknologi Sederhana di Kehidupan Sosial Masyarakat. *Jurnal Sains Dan Teknologi*, *5*(2), 157–162.

Yang, X., Xu, S., Wu, H., & Bie, R. (2019). Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network. *Procedia Computer Science*, *147*, 361–368. https://doi.org/10.1016/j.procs.2019.01.239

Zhang, X., Zhao, J., & Lecun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, *2015-Janua*, 649–657.

Zhang, Y., & Wallace, B. (2015). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. http://arxiv.org/abs/1510.03820

Zhong, B., Xing, X., Love, P., Wang, X., & Luo, H. (2019). Convolutional neural network: Deep learning-based classification of building quality problems. *Advanced Engineering Informatics*, *40*(February), 46–57. https://doi.org/10.1016/j.aei.2019.02.009