Journal of Computer Science

Original Research Paper

A Content Filtering from Spam Posts on Social Media using Weighted Multimodal Approach

Chastine Fatichah, Wildan F. Lazuardi, Dini A. Navastara, Nanik Suciati and Abdul Munif

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article history
Received: 08-11-2020
Revised: 18-01-2021
Accepted: 20-01-2021

Corresponding Author:
Chastine Fatichah
Department of Informatics,
Institut Teknologi Sepuluh
Nopember, Surabaya, Indonesia
Email: chastine@if.its.ac.id

Abstract: The system for filtering spam posts on social media is preferred to obtain the relevant content and expected by users. The previous works on spam detection have been done to filter irrelevant content on email and social media based on text or image separately. Due to the social media posts are commonly in the form of image, text, or both, the multimodal data is preferred to improve the capability of system in handling filtering content on social media. In addition, a spam post containing multimodal data sometimes does not indicate spam in both data but only one. To improve the performance of system, we propose a weighted multimodal approach for filtering content from spam posts in social media using Convolutional Neural Network (CNN). The mechanism of weighted multimodal is by weighting of spam prediction results from image and text data. We also investigate the performance of CNN architectures for spam post detection that are 3-layer, 5-layer, AlexNet and VGG16. The performance of each architectures is evaluated by 8000 Indonesian posts in the form of image and text taken from Instagram posts. The results show that the highest accuracy achieves 0.9850 based on the combination of image and text by using a 5-layer architecture. The average accuracy of all CNN architectures using multimodal data is higher than only using image and text data separately.

Keywords: Content Filtering, Spam Detection, Multimodal Data, Social Media, Convolutional Neural Network

Introduction

Spam is the use of electronic devices to transmit non-relevant messages or information to a wide number of recipients. Spam content can be found in a variety of electronic and internet media including e-mail, phone text, search engines, blog, video gaming and social media. Social media is one of technology that is currently widely used by the community as a way of exchanging information and moments in the form of message, image and videos. This social media capability allowed the irrelevant content, such as ads, to be distributed. With a large number of active users on social media allowing other people to use social media as one of the platforms to advertise their products or services. So that causes a lot of irrelevant information that is not expected by social media users. An automatic application for detecting spam in the social media to obtain the information that is useful and expected by users is preferred. Spam on social media may be in the form of comments or posts which the receiver or user does not expected.

Previous works on spam detection was conducted to filter out content in the email. The literature study of email spam filtering using the image-based filter, language-based filter, non-contents feature and collaborative spam filtering has been presented by (Blanzieri and Bryl, 2008). Wu *et al*. (2009) uses rules-based methods and neural networks to detect spam behavior on email based on text data. Previous works have been conducted on the identification of image spam email. There are two types of strategies to spam image recognition, i.e., OCR-based technique and low-level image feature technique (Biggio *et al*., 2011). Fumera *et al*. (2006) uses OCR-based techniques to extract the text embedded into images for filtering email spam. Sathiya *et al*. (2011) use combining the low-level feature and OCR-based for image spam detection. Gupta *et al*. (2012) also use the combination method of low level and metadata features for image spam on email. Aradhye *et al*. (2005) propose a spam image filtering based on extracted overlay text and color features, due to the computational expenses of

OCR-based filtering. Use edge-based features, (Nhung and Phuong, 2007) calculate the similarity score and classify the features using Support Vector Machine to detect spam image. To defeat the OCR-based approach, (Biggio *et al*., 2007) use an obscure detection approach to a low-level feature. Mehta *et al*. (2008) present two methods for image spam detection. The first approach uses visual feature and Support Vector Machine as a method of classification. The second approach is near duplicate detection of image. Annadatha and Stamp (2018) recently applied the Principle Component Analysis (PCA) and Support Vector Machine (SVM) for the identification of spam images. The Eigenspace of image spam is extracted as a feature using PCA and SVM is used to classify the image into spam or not spam. Das and Prasad (2014) proposes a combination of text extraction and low-level image feature for spam and ham categorization.

Previous works on spam detection is typically implemented in the email system and uses the conventional classification method, such as Support Vector Machine. There have been several previous studies related to spam detection on social media data. Bara (2014) propose spam detection on twitter using the likelihood approach. Zhang and Sun (2017) present user profile features and media features for spam detection on Instagram using the Random Forest method as a classification method. Research on the detection of spam on social media generally uses only text data and conventional classification methods. Recently, the classification method using deep learning approaches are popular methods for classification tasks in large data with higher accuracy than conventional classification methods. The use of deep learning as classification method to detect spam on social media has been done by (Jain and Agarwal, 2016). They use Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) methods. Nevertheless, this work has still not carried out and analyzed the effects of the experiment. The semantic CNN (SCNN) is proposed by (Jain *et al*., 2018) to detect the spam based on text. A semantic layer is added for word embedding before the convolutional layers. They evaluate the proposed method using SMS spam dataset and Twitter dataset. Jain *et al*. (2019) also used deep learning methods that are CNN and Long Short Term Neural Network (LSTM) for detecting spam in social media. They introduced semantic information of the words with WordNet and ConceptNet. They also only use the text data for spam detection.

A DeepImageSpam is proposed by (Kumar and KP, 2018) for spam detection on image spam dataset using basic CNN architecture. Fatichah *et al*. (2019) propose image spam detection using CNN method from Instagram post and compare several CNN architectures. However, the improvement of accuracy is still challenging for this research. The use of images to detect spam on social media has problems in accuracy due to they generally post message or information on social media using only text.

A content filtering based on multimodal data from spam post in social media data was not yet explored in the previous research. Due to the spam posts in social media are generally a combination of text and image, the usage of multimodal data for filtering the spam content in social media is required to improve the capability of spam detection system. In addition, a spam post containing image and text data does not indicate spam in both data sometime only in one of them. Therefore, a multimodal data is expected to improve the accuracy of spam detection.

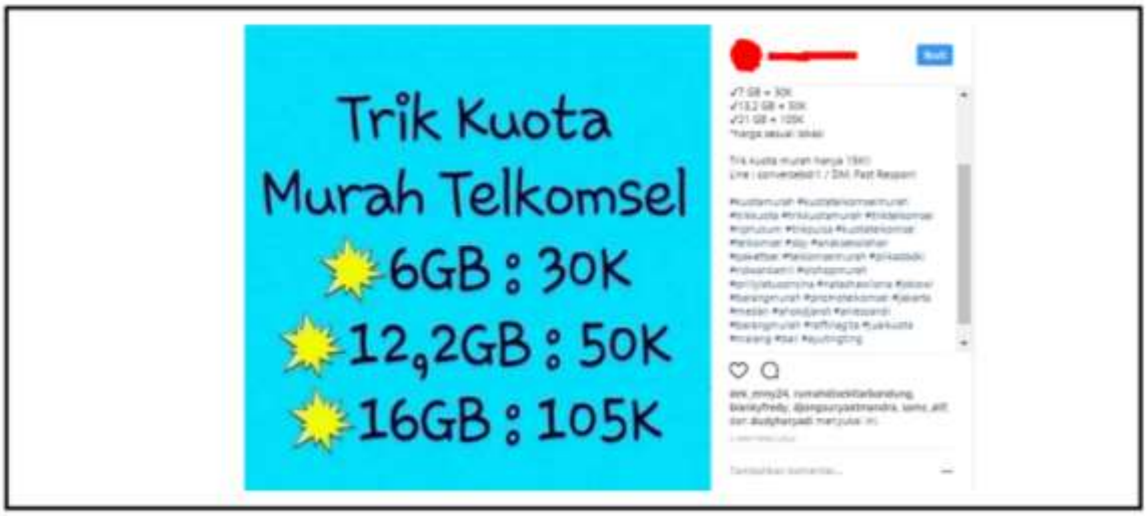
The objectives of this study are to detect the spam posts in social media using weighted multimodal approach. There are five phases of our proposed method that are crawling data, preprocessing, building the CNN model of multimodal data, testing process of multimodal data and spam detection using a weighted multimodal approach. The contribution of this research is a combining of images and text prediction results and to investigate the performance of CNN architectures in spam detection. The mechanism of combined prediction is a weighting of spam prediction results from multimodal data to obtain the final spam detection. We use CNN method due to having very good performance for many application fields and also can be used for image and text data. CNN generally consists of three types of layers i.e. convolutional layers, subsampling layers and fully connected layers. We evaluate four CNN architectures for spam detection, namely 3-layer, 5-layer, AlexNet and VGG16. The performance of each architectures is evaluated based on the accuracy.

The remainder of the paper is organized as follows: The materials and methods are presented in the section 2. The experimental results for spam detection are reported in section 3. The conclude of this research is presented in section 4.

Materials and Methods

A Spam Posts Detection on Social Media Using Weighted Multimodal Approach

The spam posts from social media is commonly in form combination image and text. The sample of spam post from social media is shown in Fig. 1. With a large number of active users on social media allowing other people to use social media to advertise their products or services. So that causes a lot of irrelevant information that is not expected by social media users. An automatic detecting spam in social media for filtering the content posts to obtain the information that is useful and expected by users is preferred.



Trik Kuota Murah Telkomsel
6GB : 30K
12,2GB : 50K
16GB : 105K

[redacted]

√7 GB = 30K
√13,2 GB = 50K
√21 GB = 105K
*harga sesuai lokasi

Trik kuota murah hanya 15K!!
Line : [illegible] / DM. Fast Respon

[illegible]

[illegible]

Tambahkan komentar...

Fig. 1: The sample of spam posts in Instagram social media

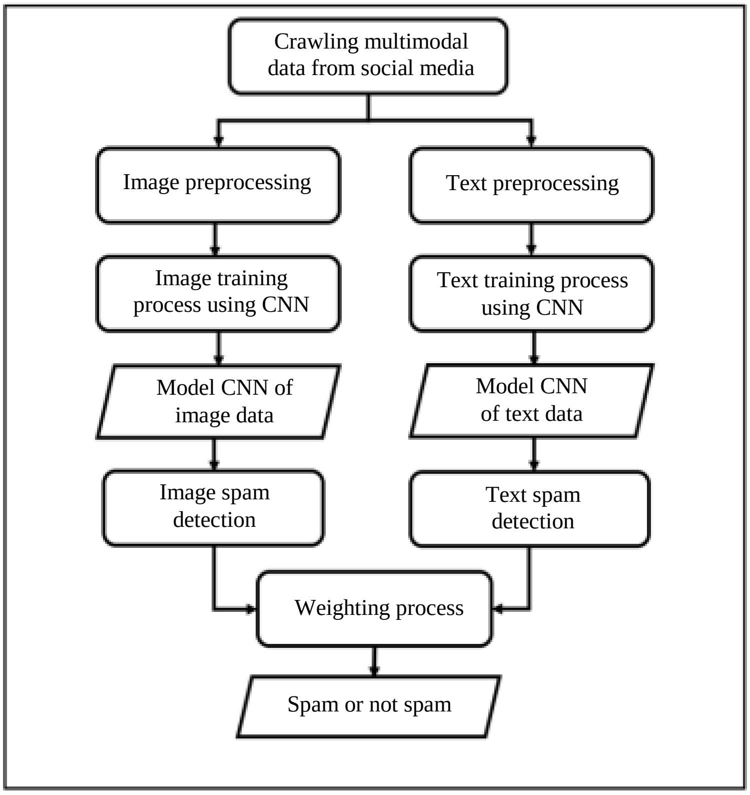


Fig. 2: The diagram system of proposed method

The proposed method for a content filtering from spam posts on social media using a weighted multimodal approach is illustrated in Fig. 2. The proposed method consists of four phases, i.e., crawling data from social media, building the Convolutional Neural Network (CNN) model of multimodal data as training process, the testing process using CNN model and the spam detection using a fusion of image and text prediction results by weighting process. The first phase is crawling data of social media that conducted through a web crawler. This research focuses on Instagram social media, so data is taken from each Instagram post in the form of image and text. The second phase performs the training process to build the CNN model. The training process is conducted for image and text data separately. The third phase conduct testing process based on CNN model. The testing process is also conducted for image and text data separately. The last phase combines the results of the testing process of image and text using the weighting process. Finally, the output of the system is the labelling of testing data that spam or not spam.

Image Preprocessing

Before carrying out the training process on image data, the image dimensions are changed to a size of 150×150 pixels. This is done because of the variant of input image sizes and also to reduce the computation process.

Text Preprocessing

The text pre-processing aims to clean the text of word particles or characters that are not needed in the next process. The cleansing process carried out in this study consists of several stages, namely:

- Eliminate website or HTML addresses
- Eliminating mention
- Remove punctuation except for fences (#)
- Eliminate numbers
- Remove symbols and emoji
- Change all letters to lowercase
- Stop words removal
- Erase words that are two letters or less

After the cleansing process is carried out, then the word embedding process is applied. This process converts the sentence into a collection of words and represents number of each word. In the word embedding process, the number of dimensions of the array must be determined in advance so that all text data that is processed has the same size. This is important because CNN requires data with a fixed dimension for processing. We use 150 dimensions of word embedding.

Training Process using Convolutional Neural Network (CNN)

The training process of image and text are done separately to build the CNN model. In this research, we compare the performance of CNN architectures such as 3-layer, 5-layer, AlexNet and VGG16.

Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is an algorithm developed from an artificial neural network that is widely used for pattern recognition and object recognition in the image. CNN is a method that combines the process of convolution as pattern analysis and the process of classification as recognition. The CNN architecture generally consists of three types of layers, namely convolutional layer, sub-sampling layer and fully connected layer (LeCun *et al*., 1990) as illustrated in Fig. 3.

The LeNet-5 model (LeCun *et al*., 1998) is the first CNN architecture for image classification with good results. The current development of CNN uses the LeNet model as the basis and then changes the size and sequence of the layers. The convolution operation is applied in the convolution layers of CNN. In mathematics, the convolution is an operation between two functions (Zhang *et al*., 2019). The notation of convolution operation is typically with an asterisk (Goodfellow *et al*., 2016) as Equation 1:

$$s(t) = (x * w)(t) \tag{1}$$

where the function *x* is an input and the function *w* is a kernel. The output *s* is called to as the feature map. The discrete convolution is defined as Equation 2 when we assume that *x* and *w* are applied only on integer *t*:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{2}$$

If two-dimensional image *I* use an input and the two-dimensional *K* use a kernel, then we obtained as Equation 3:

$$S(i,j) = (I * K)(i.j) = \sum_m \sum_n I(m,n)K(i-m,i-n) \tag{3}$$

To generate a set of linear activations, the convolutional layers perform multiple convolutions in parallel. A nonlinear activation function, such as the Rectified Linear Unit (ReLU), is used to operate each linear activation.

In the subsampling layer, a pooling function is used to adjust the output of the layer further. A pooling function substitutes a summary statistic of the nearby outputs for the net's output at a certain location. For example, the max-pooling operation produces the maximum output within a rectangular neighbourhood.

In this research, we use four CNN architectures i.e., 3-layer, 5-layer, AlexNet and VGG16 to develop the CNN model. The four CNN architectures are applied to both multimodal data i.e., image and text.

The 3-Layer Architecture

The 3-layer architecture has three layers of convolution as well as two completely connected layers and each layer of convolution ends with a pooling layer. The number of kernels of each convolutional layer are 32, 32, 64, respectively. The architecture of 3-layer is shown in Fig. 4.

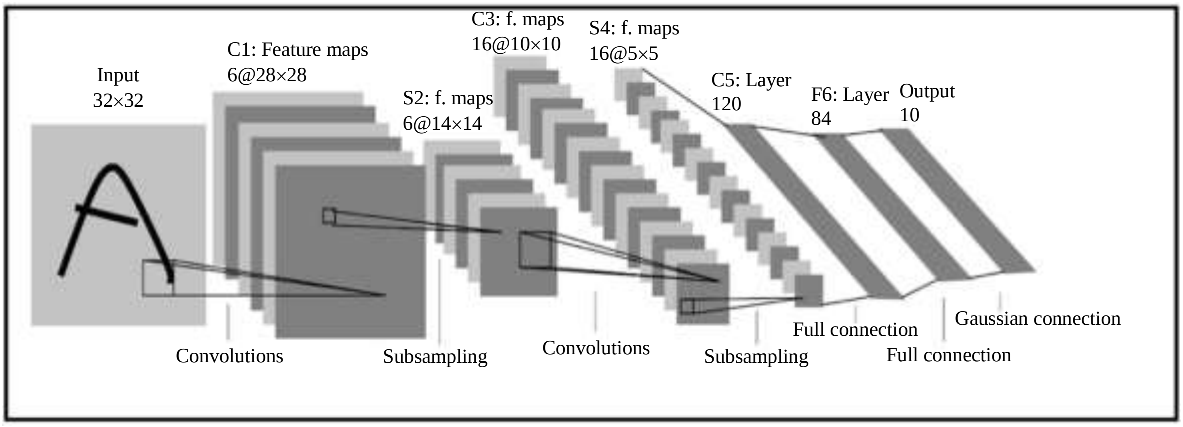


Fig. 3: The architecture of LeNet-5 (LeCun *et al*., 1998)

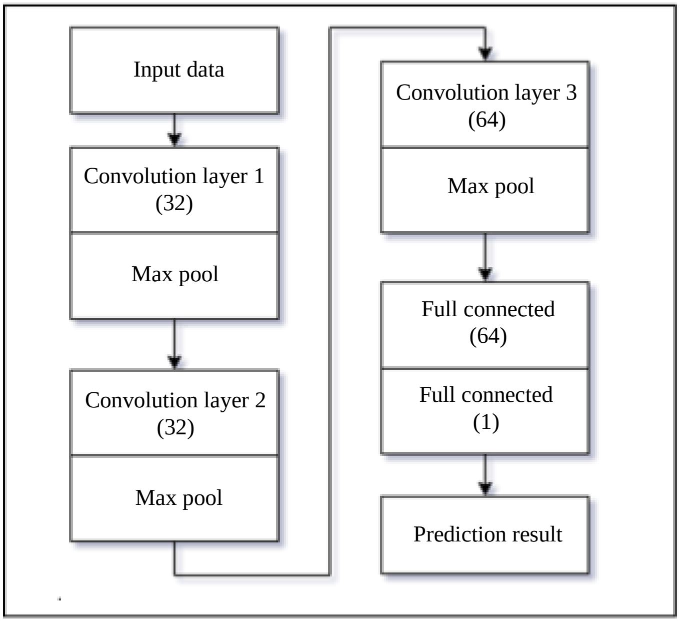


Fig. 4: The 3-layer architecture of CNN

The 5-Layer Architecture

The 5-layer architecture has five convolution layers and two fully connected layers and a pooling layer also ends with each convolution layer. The number of kernels of each convolutional layer are 96, 256, 384, 384, 256, respectively. The architecture of 5-layer is shown in Fig. 5.

The AlexNet Architecture

AlexNet is one of CNN architecture produced during the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition by (Krizhevsky *et al*., 2017). AlexNet has secured the smallest mistake of 15.4 per cent in the ILSVRC competition. It is a visual recognition achievement that deep learning can produce impressive results in accuracy. The architecture of AlexNet consists of five convolution layers and three fully connected layers. We change the number of neurons in each layer due to appropriate existing hardware capabilities in this study. We design the number of kernels of each convolutional layer are 32, 32, 64, 64, 128, respectively. The architecture of AlexNet is shown in Fig. 6.

The VGG16 Architecture

VGG16 is also one of CNN architecture that gets runners-up in the 2014 ILSVRC competition. Simonyan and Zisserman (2014) propose VGG16 architecture that consists of sixteen convolution layers and three fully connected layers. We also change the number of neurons on each layer in this study, due to the testing hardware capabilities. The architecture of VGG16 is shown in Fig. 7.

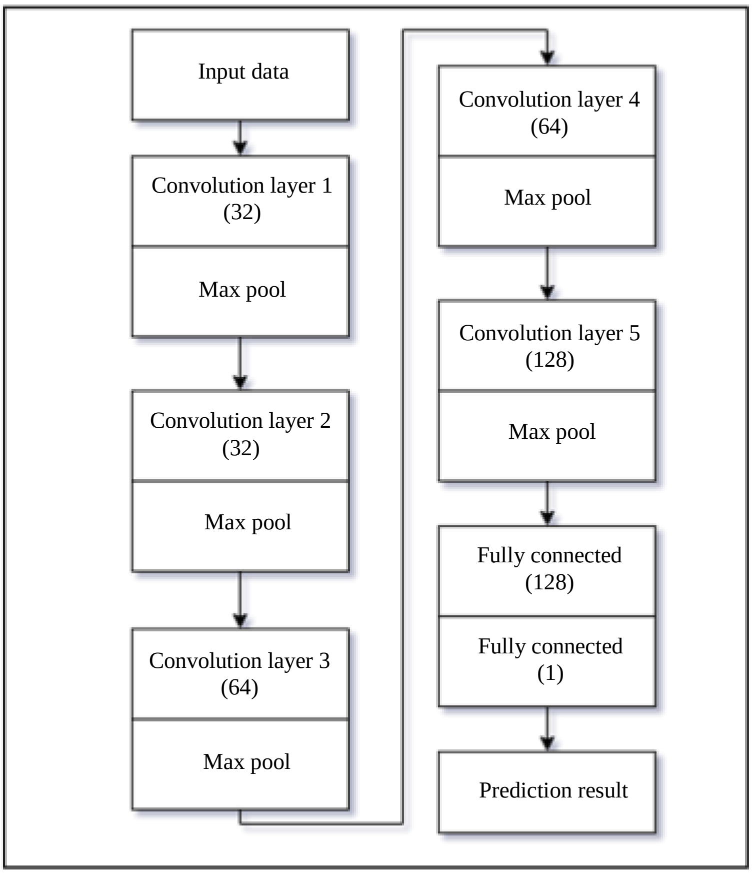


Fig. 5: The 5-layer architecture of CNN

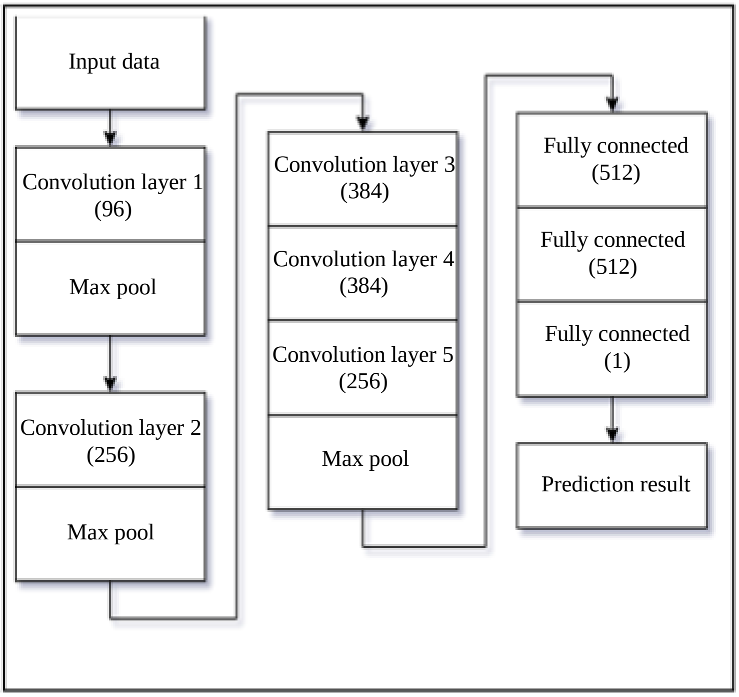


Fig. 6: The AlextNet architecture

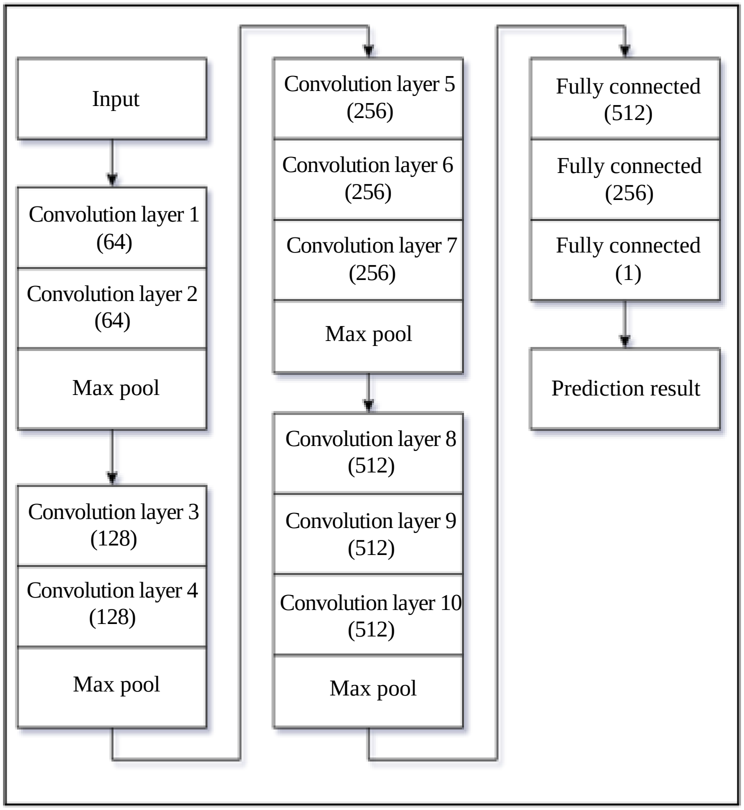

Input
Convolution layer 1 (64)
Convolution layer 2 (64)
Max pool
Convolution layer 3 (128)
Convolution layer 4 (128)
Max pool
Convolution layer 5 (256)
Convolution layer 6 (256)
Convolution layer 7 (256)
Max pool
Convolution layer 8 (512)
Convolution layer 9 (512)
Convolution layer 10 (512)
Max pool
Fully connected (512)
Fully connected (256)
Fully connected (1)
Prediction result

Fig. 7: The VGG16 architecture

The Weighted Multimodal Approach

The previous works on spam detection have been done to filter irrelevant content on social media based on text or image separately. The social media posts are commonly in the form of image, text, or both. The multimodal data is preferred to improve the capability of system in handling filtering content on social media. A spam post containing multimodal data sometimes does not indicate spam in both data but only one. For example, the first row of Table 1 show that the image post looks normal however the text post is categorized a spam. The second row of Table 1 show that both post of image and text are categorized non-spam. The use of multimodal data will consider the prediction results of both data, therefore is expected to complement each other's and to improve the performance of prediction. To improve the performance of system, we propose a weighted multimodal approach for filtering content from spam posts in social media using Convolutional Neural Network (CNN). The mechanism of weighted multimodal is by weighting of spam prediction results from image and text data.

The last phase of our proposed method is the testing process for text and image spam detection based on the CNN model to obtain the detection result. Each of CNN model predicts the class of testing data by giving a value between 0 and 1. If the value is closer to 0, the prediction result is more inclined to the normal class and the opposite applies to the spam class. The result of text spam prediction is combined with the result of image by weighting process to obtain the final spam detection. The weighting process is a fusion of the spam prediction results from image and text data. The fusion prediction (f_p) is calculated using Equation 4:

$$f_b = w_i p_i + w_t p_t \tag{4}$$

where, p_i is the prediction result of image spam detection, p_t is the prediction result of text spam detection, w_i is the weight of image prediction and w_t is the weight of text prediction. If the f_p value is less or equal to 0.5, the testing data is classified as a not spam, otherwise, if the f_p value is more than 0.5 then the testing data is classified as spam.

Experiments on Spam Posts from Social Media

Dataset Description

We use 8000 Indonesian posts (Fatichah *et al*., 2019) taken from Instagram and is divided into 7600 as training data and 400 as testing data. The posts are in the form of image and text. The testing data consist of two classes i.e., spam and normal with 200 data for each class. The example of spam post and normal post on Instagram are shown in Table 1.

Experiments on Spam Detection based on Multimodal

This section explains the scenario of experiments that have been carried out on spam detection as follows:

Scenario 1: Determining the best weighting values for the multimodal posts.
Scenario 2: Calculating the accuracy values on the 3-layer model based on the kernel dimension variation.
Scenario 3: Calculating the accuracy values on the 5-layer model based on kernel dimension variation.
Scenario 4: Comparing of the accuracy values on 3-layer, 5-layer, AlexNet and VGG16 architectures

The training process in each scenario has several fixed parameters as shown in Table 2.

The fusion result (f_p) is calculated using equation 4. The weight values of w_i and w_t are determined in scenario 1. If the f_p value is less or equal to 0.5, the testing data is classified as a normal class; otherwise, if the f_p value is more than 0.5 then the testing data is classified as spam. The first scenario in this experiment aims to determine the weight value for each data type. The prediction results of both types of data are multiplied by the weight then summed to get the final predictive value. There are nine variations of weight values in both data types in the range 0.1-0.9 with an interval of 0.1. We design the sum of both weights should be 1, due to the prediction result of each CNN classifiers are the probability values range in [0,1]. Therefore, the final prediction of multimodal data is also probability value range in [0,1].

Table 1: The example of Indonesian Instagram post with spam and normal label

Image post	Text post	Label
	Jam tangan Gshock Harga 185rb Kontak via whatsapp Line Rojo.jamtangan Spek. Dualtime Backlight Date alarm stopwaatch Bahan rubber Water resis skala kecil #jualjamtangan #gshockwarrior #jamgshock #jamtanganmurah #jualjamtanganmurah #jamtangan #jamtangansport #jamtangankaret #ordersekarang #jamtangankw #jamtangancowo #jam #jualgshock suka #like4like #fff #follow4follow # #depok#bekasi #mataram #bandung #palembang #batam	Spam
	Ada yg mau penghasilan nya kayak gini?? Info wa #pekanbaru #payakumbuh	Spam
	Cantiknya bunga krisan :)	Non-Spam

Table 2: The parameter of the CNN model

Parameter CNN	Image	Text
Epoch	50	10
Pool size	2×2	2×1
Dropout rate	0.5	0.5
Embedding dimension	-	150
Sequence length	-	300
Optimizer	Adam	Adam

Scenario 1 uses the AlexNet model as a test model. The combination of weight values with the best accuracy will be used in subsequent scenarios. The experimental results of scenario 1 are shown in Table 3. The results show that the highest accuracy is achieved when the image weight value is 0.4 and the text weight value is 0.6. Scenario 2 aims to calculate the accuracy values in the 3-layer architecture with kernel dimension variations. The experimental results of scenario 2 are shown in Table 4. Scenario 3 aims to calculate the accuracy values in the 5-layer architecture with kernel dimension variations. The experimental results of scenario 3 are shown in Table 5. The highest accuracy of image, text and multimodal data are achieved by 5-layer with kernel dimension 2×2. Scenario 4 aims to compare the accuracy values of the 3-layer, 5-layer, AlexNet and VGG16 architectures. The parameter of CNN architectures such as the number of layers and number of neurons are defined in Table 6.

The experimental results of scenario 4 are shown in Table 7. The highest accuracy of image data is obtained by VGG16 architecture and the accuracy is 0.8475. The highest accuracy of text and fusion data is achieved by the 5-layer architecture with kernel dimension 2×2 and the accuracy is 0.9850. The average accuracy of all CNN architectures using fusion data is 0.9775 and is higher than only using image and text are 0.7675 and 0.9731, respectively. The 5-layer architecture also achieves the lowest of training time in image spam detection. But, the lowest of training time in text spam detection is obtained by the 3-layer architecture. The VGG16 have the highest training time of both image and text data.

The confusion matrix of prediction results with the highest accuracy by 5-layer architecture using a multimodal data is shown in Table 8. The prediction results show that the most misclassification when the spam categories are predicted as the non-spam class. The misclassification results are commonly image spam that predicted as non-spam image. The prediction results show that the most correct classification when the non-spam categories are predicted as the non-spam class.

Table 3: The spam detection results using weighted multimodal

Values of image weight (*wi*)	Values of text weight (*wt*)	Accuracy
0.9	0.1	0.7550
0.8	0.2	0.7650
0.7	0.3	0.7825
0.6	0.4	0.8050
0.5	0.5	0.9375
0.4	**0.6**	**0.9825**
0.3	0.7	0.9775
0.2	0.8	0.9775
0.1	0.9	0.9750

The example of the correct prediction results for spam detection is shown in Table 9. The first row in Table 9 is a sample with actual non-spam class and the prediction class is also non-spam. The second and third row is a sample with actual spam class and the prediction class is also spam. The example of the incorrect prediction results for image spam detection is shown in Table 10. There are some spam images that ads woman dress in incorrect prediction due to the many image samples of training data for the non-spam category is the woman dress. In contrast, the example results in Table 10 is misclassification because of scant image samples that similar to image example on spam categories.

Table 4: The accuracy of 3-layer architecture with kernel dimension variations

	Accuracy		
Kernel dimension	Image	Text	Multimodal
2×2	0.6975	0.9650	0.9675
3×3	0.7025	0.9775	0.9800
4×4	0.7400	0.9824	**0.9825**
5×5	0.7175	0.9425	0.9475

Table 5: The accuracy of 5-layer architecture with kernel dimension variations

	Accuracy		
Kernel dimension	Image	Text	Multimodal
2×2	0.7375	0.9825	**0.9850**
3×3	0.7025	0.9775	0.9775
4×4	0.7250	0.9700	0.9800
54333[a]	0.7300	0.9600	0.9675

a. The kernel dimension variation in each layer from the first layer in sequence by 5×5, 4×4, 3×3, 3×3 and 3×3

Table 6: The parameter of CNN architectures

		No of Neurons	
Architecture of CNN	No of layers	Image	Text
3-layer (4×4)	8	193	193
5-layer (2×2)	12	449	449
AlexNet	11	2401	1201
VGG16	17	4417	4993

Table 7: The comparison of accuracy values in 3-layer, 5-layer, AlexNet and VGG16 architectures

	Accuracy		
Architecture of CNN	Image	Text	Multimodal
3-layer (4×4)	0.7400	0.9824	0.9825
5-layer (2×2)	0.7375	0.9825	**0.9850**
AlexNet	0.7550	0.9750	0.9825
VGG16	0.8475	0.9525	0.9600
Average	**0.7700**	**0.9731**	**0.9775**

Table 8: The confusion matrix of prediction results by 5-layer architecture with kernel dimension 2×2 using a multimodal data

		Prediction		
		Normal	Spam	No. of each category
Actual	Normal	200	0	200
	Spam	6	194	200
		Total of testing data		400

Table 9: The example of correct prediction results by 5-layer architecture with kernel dimension 2×2 based on multimodal

Image post	Text post	Actual class	Prediction class
	Cantiknya bunga krisan :)	Non-Spam	Non-Spam
	Ada yg mau penghasilan nya kayak gini?? Info wa #pekanbaru #payakumbuh	Spam	Spam
	Dress midi tie dye ekslusif Dengan bahan jersey yang sangat nyaman dipakai, tekstur halus Ukuran fit to XL Ld 110 cm bahan melar hingga ld 120 Panjang 105 cm Motif tie dye yang unik Harga ecer 55.000 Seri: 50.000/pcs Open reseller Menerima orderan puluhan, ratusan hingga ribuan #medan #tangerang #surakarta #jakarta #DKI #makassar #malang #lumajang #samarinda #batam #palembang #padang #balikpapan #mataram #dressmidi		

Table 10: The example of incorrect prediction results for image spam detection

Image post	Actual class	Prediction class
	Spam	Non-Spam

Table 11: The example of incorrect prediction results for text spam detection

Text post	Actual class	Prediction class
Key Holder/dompet STNK - GC Rp.25.000 #keychain #keyholder #dompetstnk #dompetkunci #dompetkuncimobil #dompetkuncimotor #kuncimobil #gucireplika #kuncimotor #leatherkeychain #brown #coklat #bogor #olshopbogor #indie #indiesthings	Spam	Non-Spam

The example of the incorrect prediction results for text spam detection is shown in Table 11. The first example in Table 11 is wallet advertising. The wrong prediction is probably due to the rarity of the wallet advertising on training data and the lack of terms that refer to the spam category. While the second example is a promotion with many mentions and hashtags. The incorrect prediction occurs due to the term in the form of mention is not considered in the training process because all of mentions have been removed during the pre-processing text.

Conclusion

A spam post containing multimodal data does not indicate spam in both data sometimes only in one of them. So, multimodal data is expected to improve the accuracy of spam detection. This research focuses on spam detection from social media posts based on a weighted multimodal approach. We also investigate the performance of CNN architectures for spam detection that are 3-layer, 5-layer, AlexNet and VGG16. The performance of each architecture is evaluated by Indonesian posts in the form of images and text taken from Instagram. The highest accuracy of image data is obtained by VGG16 architecture and the accuracy is 0.8475. The highest accuracy of multimodal data is achieved by the 5-layer architecture with kernel dimension 2×2 and the accuracy is 0.9850. The average accuracy of all CNN architectures using multimodal data is 0.9775 and is higher than only using image and text are 0.7675 and 0.9731, respectively. The final spam detection results are calculated by a combination of image and text prediction results with the weight values of 0.4 for image and 0.6 for text.

The proposed method can be applied to other social media data. However, the collecting of sample social media posts for the training process should be increased to improve the performance results. In addition, the other CNN architectures or deep learning methods can be used to measure the best performance of architectures. In future research, multimodal data can be extended for incident detection in social media such as emergency incident.

Acknowledgement

This research is funded by the Indonesian Ministry of Research Technology through PTUPT grant.

Author’s Contributions

Chastine Fatichah: Contributed to define the research problem, formulation of methods and writing the paper.

Wildan F Lazuardi: Assisted in the implementation of methods and conduct of the experiment.

Dini A. Navastara, Nanik Suciati and Abdul Munif: Assisted in writing the paper.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Annadatha, A., & Stamp, M. (2018). Image spam analysis and detection. Journal of Computer Virology and Hacking Techniques, 14(1), 39-52.

Aradhye, H. B., Myers, G. K., & Herson, J. A. (2005, August). Image analysis for efficient categorization of image-based spam e-mail. In Eighth International Conference on Document Analysis and Recognition (ICDAR'05) (pp. 914-918). IEEE.

Bara, I. A. (2014). Discovering Spam On Twitter.

Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2007, August). Image Spam Filtering by Content Obscuring Detection. In CEAS.

Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2011). A survey and experimental evaluation of image spam filtering techniques. Pattern recognition letters, 32(10), 1436-1446.

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1), 63-92.

Das, M., & Prasad, V. (2014, June). Analysis of an image spam in email based on content analysis. In Proc. Int. Conf. On Natural Language Processing And Cognitive Computing (Vol. 201, No. 4).

Fatichah, C., Lazuardi, W. F., Navastara, D. A., Suciati, N., & Munif, A. (2019). Image spam detection on instagram using convolutional neural network. In Intelligent and Interactive Computing (pp. 295-303). Springer, Singapore.

Fumera, G., Pillai, I., & Roli, F. (2006). Spam filtering based on the analysis of text information embedded into images. Journal of Machine Learning Research, 7(Dec), 2699-2720.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

Gupta, A., Singhal, C., & Aggarwal, S. (2012). Identification of image spam by using low level & metadata features. International Journal of Network Security & ITS Applications, 4(2), 163.

Jain, G., & Agarwal, B. (2016). An overview of RNN and CNN techniques for spam detection in social media. IJARCSSE, 6(10), 126-132.

Jain, G., Sharma, M., & Agarwal, B. (2018). Spam detection on social media using semantic convolutional neural network. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 8(1), 12-26.

Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short term memory neural network. Annals of Mathematics and Artificial Intelligence, 85(1), 21-44.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

Kumar, A. D., & KP, S. (2018). Deepimagespam: Deep learning based image spam detection. arXiv preprint arXiv:1810.03977.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems (pp. 396-404).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Mehta, B., Nangia, S., Gupta, M., & Nejdl, W. (2008, April). Detecting image spam using visual features and near duplicate detection. In Proceedings of the 17th international conference on World Wide Web (pp. 497-506).

Nhung, N. P., & Phuong, T. M. (2007, March). An efficient method for filtering image-based spam. In 2007 IEEE International Conference on Research, Innovation and Vision for the Future (pp. 96-102). IEEE.

Sathiya, V., Divakar, M., & Sumi, T. S. (2011). Partial image spam E-mail detection using OCR. Int J Eng Trends Technol, 1(1), 55-59.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Wu, C. H. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert systems with Applications, 36(3), 4321-4330.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2019). Dive into deep learning. Unpublished Draft. Retrieved, 19, 2019.

Zhang, W., & Sun, H. M. (2017, January). Instagram spam detection. In 2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 227-228). IEEE.