Sentiment Analysis of COVID-19 using Multimodal Fusion Neural Networks

Ermatita Ermatita, Abdiansah Abdiansah, Dian Palupi Rini, Fatmalina Febry

Universitas Sriwijaya, Palembang, Indonesia

Abstract **– The purpose of this study creates a Sentiment Analysis model of COVID-19 using Multimodal Fusion Neural Networks in real time to model and visualize COVID-19 in Indonesia. This study obtained 87 percent accuracy using the Multimodal Fusion Neural Networks model, a higher 5 percent than the benchmarking model Convolutional Neural Networks. This study proves that the sentiment model built is quite promising and relevant to be implemented.**

Keywords **– Multimodal Fusion Neural Networks, COVID-19, Sentiment Analysis.**

1. Introduction

Sentiment Analysis is an approach that focuses on analyzing sentiment from a particular domain such as text or images. Most of the sentiment analysis models currently used focus on using Twitter data for classification. Ye et al. [1] approach to reviewing tourist destinations using machine learning. In Anastasia et al. [2] sentiment analysis was used to show the level of satisfaction of transportation consumers online. In addition to Twitter data, some researchers also use advanced models for sentiment analysis of texts such as [3], [4].

DOI: 10.18421/TEM113-41
https://doi.org/10.18421/TEM113-41

Corresponding author: Ermatita Ermatita,
Universitas Sriwijaya, Palembang, Indonesia
Email: ermatita@unsri.ac.id

Received: 29 March 2022.
Revised: 10 August 2022.
Accepted: 16 August 2022.
Published: 29 August 2022.

Aspect-based sentiment analysis has become one of the expansions in sentiment analysis research with the emergence of various models such as Schmitt et al. [5] which uses end-to-end Convolutional Neural Networks CNN as the main architecture in aspect-based sentiment analysis; Li et al. [6] which utilizes the pre-trained BERT model as the main architecture in aspect-based sentiment analysis and so on. The use of state-of-the-art models of natural language processing such as BERT [7] which is widely used as a pre-trained model, can help in improving the evaluation performance of the developed model, but challenges such as if the data is composed of more than one modality such as an image will affect the quality of the model. Based on this, the first challenge is how to make a model that can accept input in the form of multimodal features in the form of text captions and images for the case of sentiment analysis. Some models that already support multimodal, such as hierarchical-level Fusion [8], can accept input data in the form of a bottleneck layer. Such as comes from the text and voice features but is still limited in speed performance in inference if the input is in the form of images and text. Other models such as Graph Fusion Encoder [9] which is currently state-of-the-art in the machine translation domain with multimodal features can combine images and parallel text corpus to produce good model performance. Currently, the model is still used in the machine translation domain only. This model can be adopted and modified to suit the case of sentiment analysis with multimodal feature images and text captions. Utilization of the latest technology must be balanced with training time and fast inference which also depends on the complexity of the model used, for example, some architectures with CNN networks background [10] will have an O(n/k) model complexity while some other architectures are based on Long Short-Term memory [11] will have a slower O(n) complexity than CNN. The second challenge is how to create a sentiment analysis model that can accept input in the form of multimodal features that can perform training and inference quickly and with low complexity.

Currently, the COVID-19 case in Indonesia is very high, with more than one million cases occurring until March 2021. The use of sentiment analysis in analyzing a community's mood towards COVID-19 is one way to find out public sentiment regarding the handling and prevention of COVID-19 through data on social media such as Instagram. For example, if we want to observe the success or public sentiment of implementing vaccines in Indonesia based on data from social media such as Instagram, the multimodal sentiment analysis model can be used. Therefore, the third challenge is how to use the multimodal sentiment analysis model, feature images, and text captions to be able to observe public sentiment on the COVID-19 case. The fourth challenge is how to collect datasets from Instagram related to caption texts and post images regarding the prevention and control of COVID-19 in Indonesia. The contributions of this research are as follows: (1) the creation of a novelty model that can handle input in the form of multimodal features that can analyze sentiment, (2) collection of a multimodal dataset of image features and captions from Instagram, (3) the use of the model that has been made in the sentiment case community regarding the handling and prevention of COVID-19 in Indonesia.

Based on the above explanation, in this study we build a Sentiment Analysis model for COVID-19 using Multimodal Fusion Neural Networks in real-time. We face these problems: (1) How to create a model that can handle input in the form of multimodal feature images and text for the case of sentiment analysis, (2) Collection of multimodal feature datasets from Instagram sources in the form of post images and caption texts with the topic of COVID-19, (3) How to implement the model that has been built in the case of COVID-19. The purpose of this research is as follows: (1) To create a model that can handle input in the form of multimodal feature images and text for the case of sentiment analysis, (2) To collect a multimodal feature dataset from Instagram sources in the form of post images and caption texts with the topic of COVID-19, (3) To implement the model that has been built in the case of COVID-19.

2. Related Works

In the focus of the Sentiment Analysis is analyzing text data. In its use, most research uses data of Twitter to collect sentiment for the classification process. Sonia and Indra [2] perform sentiment analysis research to show which online transportation Grab and GO-JEK are more satisfying to users. In their study, it has been proven that the NSS calculation has the same results as the traditional customer satisfaction method.

Deep Neural Networks is a Deep Learning architecture that is composed of more than one hidden layer that accepts input data to generate output predictions (Figure 1.) [12]. This architecture attempts to match the prediction to the label (Ground Truth) by tuning the parameters.

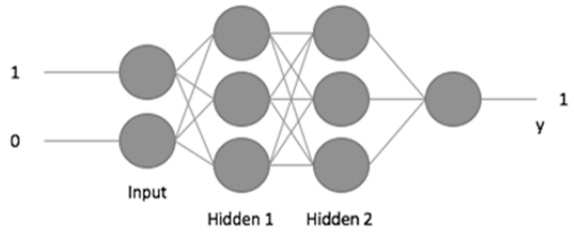

Figure 1. Deep Learning Architecture

The input in this architecture consists of (*input layer*)$x = (x_1, x_2, x_3 \ldots, x_T)^T$, (*hidden layer*) $h = (h_1, h_2, h_3 \ldots, h_T)^T$ and (*Output layer*) $y = (y_1, y_2, y_3 \ldots, y_T)^T$. Then there is the weight matrix W= {W_ij} which connects the input layer with the hidden layer and the weight matrix U= {U_ij} which connects the hidden layer h_i with the output layer r_i. The calculation of the hidden layer in this architecture is as follows:

$$s_j = \sum_{i=1}^{n} x_i W_{ij} \quad (1)$$

$$r_j = \sum_{i=1}^{n} h_i U_{ij} \quad (2)$$

The output of the hidden layers s_j (Eq.1) and r_j (Eq.2) will each be assigned an activation layer to transform the linear into non-linear functions as follows:

$$h_j = f(s_j) \quad (3)$$

$$y_j = f(r_j) \quad (4)$$

Some of the activation functions that are often used are Tangent Hyperbolic (tanH), Logistics Function (sigmoid), and Rectified Linear Unit (ReLU) shown in Figure 2.:

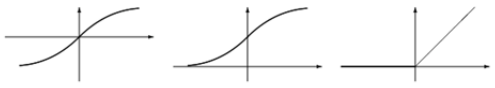

Figure 2. The activation functions

$$tanH = \frac{2}{1+e^{-2x}} - 1 \quad (5)$$

$$Sigmoid = \frac{1}{1+e^{-x}} \quad (6)$$

$$ReLU = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \quad (7)$$

The output of y_j (prediction result) will be compared with y_o (output / Ground Truth), then the loss function is calculated with Mean Square Error (MSE). Gradient Descent will then work on training data and calculate the Gradient of each previous layer to change the weight (Weight). This continues until the Convergence or weight does not change anymore.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - y_o)^2 \quad (8)$$

Convolutional Neural Networks are the same as the DNN architecture discussed previously, which consists of neurons that learn weights and biases, each neuron receives input and performs matrix multiplication and then enters non-linearity such as ReLU. Networks still have a loss function and a fully-connected layer. In addition, this architecture is equipped with a convolutional layer and a max-pooling layer as shown below.

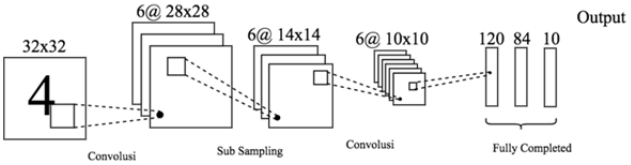

Figure 3. The activation functions

In Figure 3., for example, the input image x_ijk 32x32x3 pixels will enter the convolution layer S(x,W) where the filter or kernel W_ij 5x5 will work to perform the convolution or scanning process on receptive fields to produce 6 layer feature maps 28x28.

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_{i=1}^{n}\exp(z_j)} \quad (9)$$

Recurrent Neural Networks are similar to the DNN architectures discussed above. The difference is that this architecture uses the value of the previous state (previous state) to calculate the current state. The RNN has the ability to link previous information to the current task, for example using the previous video frame will provide insight into the current frame (Figure 4.).

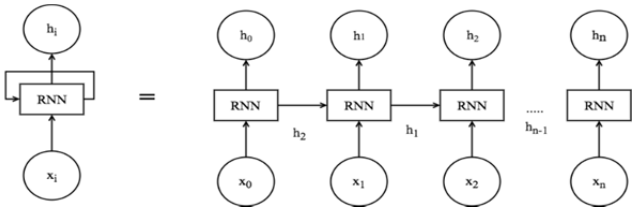

Figure 4. RNN

The weakness of the RNN is that the previous state is always used in the current state which is often referred to as "long term dependency" which is often not very important to use. This is what underlies the emergence of variations of the RNN, namely LSTM [11]. The LSTM is designed to avoid long-term dependencies, namely remembering information for a long time. The complete architecture of the LSTM can be seen in Figure 5:

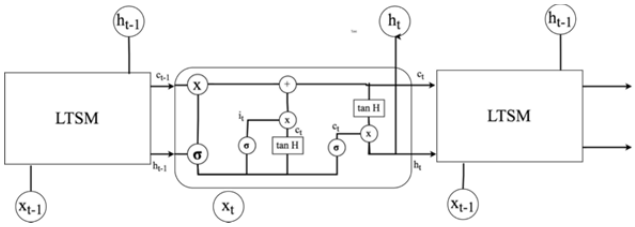

Figure 5. LSTM architecture

In the picture above, it can be seen that the LSTM architecture is the same as the RNN, which is to get the previous cell state value but has several additional gates. The first step of the LSTM is to determine whether the information from the previous cell state is forgotten or stored in the current cell with the forget gate f_t:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (10)$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (11)$$
$$c_t = tanH(W_c[h_{t-1}, x_t] + b_c) \quad (12)$$
$$C_t = f_t * C_{t-1} + c_t \quad (13)$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (14)$$
$$h_t = o_t * tanH(C_t) \quad (15)$$

The next step is to determine what information will be stored in the cell state with the help of 2 parts: (1) the input gate i_t which determines what value to update and (2) the new candidate c_t. The calculated candidate c_t will be multiplied by the input gate i_t to prepare it to be the current state c_t= i_(t)* c_t. The previous state C_(t-1) is then determined whether to be used in the current state by multiplying by f_t and adding up the candidate c_t. The last step is to determine what value to output by calculating the gate output o_t. The current hidden layer h_t is calculated by Eq.10.

3. Methodology

3.1. Design Experiment

This research is divided into 2 stages, namely the creation of a Sentiment Analysis model that can receive input in the form of text and posters from Instagram social media. Modeling involves Deep Learning architecture with modifications to networks such as adding a multimodal Graph layer containing captions and images, adding self-attention and using fusion. All these theories have been discussed in the previous chapter. After the network is created, an evaluation will be run to measure the performance of the model using Precision, Recall and F-score. The model that has been evaluated will be used in the next stage, namely the implementation of the model in the case of COVID-19. The second stage will begin with collecting datasets from Instagram by involving crawling techniques to retrieve Instagram poster and text data. The data that has been collected will be given a sentiment in the form of a sentiment

class (positive, negative and neutral) and a mood class (happy, sad, angry and relaxed). After being given a sentiment and mood class, the dataset will be evaluated using the Kappa Score to assess the agreement between two or more annotators. The dataset that has been collected is called a corpus, where this corpus will be used to train the model. After getting a model with Instagram data related to COVID-19, the model will be evaluated for performance and run in real-time architecturally by utilizing other technologies such as Kafka and streaming processes, so that in the future this system will be able to be used to monitor real-time data from social media.

3.2. Fusion

Merging information from various features such as sound, images and text can be done by fusion techniques: (1) feature-level fusion, (2) decision-level fusion and (3) hierarchical-level fusion [8]. The last name is the state-of-the-art on the topic of spoken-dialogue which combines acoustic and lexical features with an accuracy of 57.3 percent. Hierarchical-level fusion receives input from different features and then in one of the first hidden layers a feature-level process is carried out (combining features before the recognition process is carried out) and in the second hidden layer a decision-level process is carried out (determining the final decision to determine the output). The complete architecture is shown in Figure 6.:

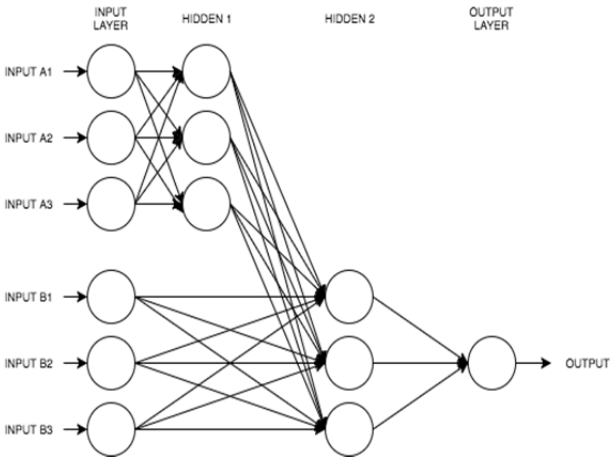

Figure 6. Hierarchical-level fusion

The fusion architecture can accept input in the form of bottleneck features [13], [14], features generated by hidden layers in neural networks with various models such as DNN, CNN, RNN and CRNN as done by [15] which combines the CNN input of acoustic features with DNN input of acoustic features. The following is a proposed model that can accept input in the form of text and images for the case of Sentiment Analysis. The model comes from a combination of several empirical models that are formed to receive input in the form of images and text from Instagram. Instagram text captions and posters can be combined with a multimodal graph which will become a unified vector that can be forwarded to visual self-attention and text self-attention. Each output of self-attention will be entered in Hierarchical Fusion.

3.3. Sentiment Analysis

The dataset is collected by taking based on keywords, namely certain in COVID-19. The collection is carried out from April 2021 - May 2021 in Indonesian. Data is collected through Instagram by taking posts, captions and sentiments that classifies into three categories, there are negative (-1), positive (1), and neutral, or incomprehensible (0). There are some rules before classifying dataset for each tweet. For the data labeling process, the researcher used 4 annotators, all of whom were students of computer science, mathematics, and computer engineering to measure the Kappa value. The reliability test of two or more annotators use Kappa value [16]. Ensuring the data using this technique to know the study has an appropriate representation is quite important. The annotator will manually label all tweet data. The labeling process is carried out according to the rules described above. Once the labeling process is complete, the sentiments of all annotator from the new dataset created will be compared. The new dataset will calculate the kappa value for each annotator. After getting the kappa value for each annotator then we get the average value to see if the kappa is considered good, moderate, or bad. As previously mentioned, this study uses two classes, namely positive (1) and negative (0). The rest considered neutral or irrelevant were not used in this study. Before the dataset can be processed into the model, the data must first be processed into a pre-processing process to remove irrelevant words that do not represent any sentiment such as punctuation, symbols, numbers, and links. The researcher also performed stemming [17], normalization [18], and removing stop words as well.

1. Remove Punctuation, like comma, period, exclamation mark, etc. must be deleted.
2. Deleting symbols and numbers.
3. Deleting a link, all tweets that contain a tweet link, the link must be deleted
4. Stemming, words in the Indonesian Dictionary KBBI that are not standardized will be changed to standard words.
5. Normalization, abbreviations will be changed to ordinary words. The dictionary will be made that containing abbreviated words and original words. For example, “very much”, “no”, etc.

6. Stop words, to remove words that do not contain sentiment in tweets. For example, “which”, “there”, “follow”, “then”. All words containing stop words will be deleted.

4. Results and Discussion

The dataset was taken from several sources, particularly on Twitter and the crawling websites, the main keyword about COVID-19, Vaccine, and others. After collecting the data, we divided it into three classes: positive, negative, and neutral, the dataset was then evaluated by using the Kappa score as follows. We obtained more than 0.7 which indicated our dataset is valid.

$$p_A = \frac{330+3}{337} = 0.981$$

$$p_{relevan} = \frac{330+2}{330+3+2+2} \cdot \frac{330+2}{330+3+2+2} = 0.970$$

$$p_{tidak} = \frac{3+2}{330+3+2+2} \cdot \frac{3+2}{330+3+2+2} = 0.0002$$

$$p_e = 0.970^2 + 0.02^2 = 0.9764$$

$$Kappa = \frac{(0.981-0.9764)}{(1-0.9764)} = 0.812$$

The dataset then changed into word2vec and BoW features as follow:

[71,147,233,128,383,201,9,616,160,536,17,54] 0
[24,9,975,79,204,641,11,10,29,475,204,15,80] 1 ...
[276,19,929,281,108,7,175,172,5,33,25,54,1] 150

The final dataset is then grouped into a CSV file which consists of the BoW features and Word2Vec features.

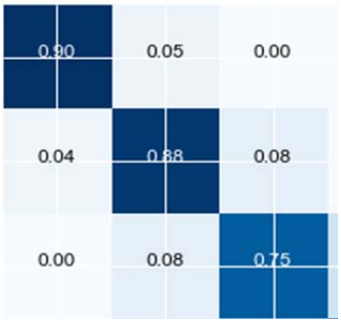

Figure 7. Confusion Matrix of Fusion

Text classification based on the LSTM model is performed by BoW and Word2Vec in order to prove and analyze the feasibility of the COVID-19 dataset that was created before. BoW and Word2Vec feature x_i will be forwarded into LSTM architecture which consists of 4 gates: input i_t in Equation 16, output o_t in Equation 17, forget f_t in Equation 18, and Candidate C_t in Equation 19. For each gate, the formula can be computed as follows:

$$i_t = \sigma(W_i * [C_{t-1}, h_{t-1}, x_t] + b_i) \quad (16)$$
$$o_t = \sigma(W_o * [C_t, h_{t-1}, x_t] + b_o) \quad (17)$$
$$f_t = \sigma(W_f * [C_{t-1}, h_{t-1}, x_t] + b_f) \quad (18)$$
$$C_t = f_t * C_{t-1} + (1 - f_t) * \sim C_t \quad (19)$$

Input gate layer decides what value will be updated, forget gate layer decides the previous state will be kept or thrown. Output gate layer decides the output and the candidate is the value that will be added into the state. Detailed LSTM architecture is shown in Figure 8.:

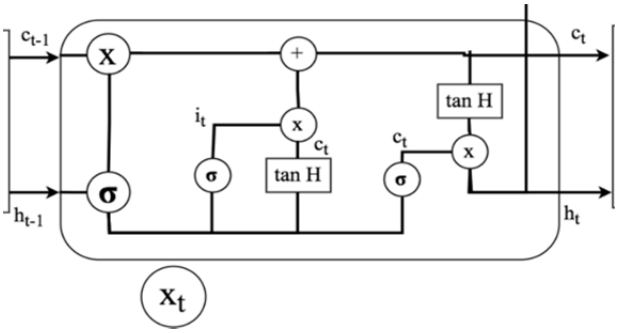

Figure 8. Long Short Term Memory

Training is performed in GPU computer GTX 1050 with CUDA Architecture in order to decrease training time. 30 epoch training was performed for each BoW and Word2Vec data, then Validation data was used to measure accuracy and loss compared with Training data in order to prevent over fitting. After performing training for each design experiment in order to produce appropriate models. Based on confusion matrix (Figure 7.), testing data is used to test the overall accuracy of the model. Complete prediction results for each design experiment can be seen in Table 1.

Table 1. Text classification result

Class	*Genre*	*Precision*	*Recall*	*F-score*	*Accuracy*
Sentiment	+	0.94	0.92	0.93	95%
	-	0.95	0.98	0.96	
	#	0.94	0.90	0.92	
Mood	Sad	0.85	0.83	0.84	85%
	Happy	0.84	0.85	0.83	
	angry	0.84	0.80	0.82	
	relax	0.84	0.80	0.82	

5. Conclusion

This study obtained 87 percent accuracy using the Multimodal Fusion Neural Networks model, a higher 5 percent than the benchmarking model Convolutional Neural Networks. This study proves that the sentiment model built is quite promising and relevant to be implemented.

References

[1]. Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, *36*(3), 6527-6535. doi:10.1016/j.eswa.2008.07.035

[2]. Anastasia, S., & Budi, I. (2016, October). Twitter sentiment analysis of online transportation service providers. In *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 359-365). IEEE. doi: 10.1109/ICACSIS.2016.7872807

[3]. Liao, W., Zeng, B., Liu, J., Wei, P., Cheng, X., & Zhang, W. (2021). Multi-level graph neural network for text sentiment analysis. *Computers & Electrical Engineering*, *92*, 107096. doi:10.1016/j.compeleceng.2021.107096

[4]. Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Tanvir Afzal, M. (2020). Important citation Identification using Sentiment Analysis of In-text citations. *Telematics and Informatics*, 101492. doi:10.1016/j.tele.2020.101492

[5]. Schmitt, M., Steinheber, S., Schreiber, K., & Roth, B. (2018). Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 1109–1114. doi: 10.5282/ubm/epub.61858

[6]. Li, X., Bing, L., Zhang, W., & Lam, W. (2019, November). Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)* (pp. 34-41). doi:10.18653/v1/D19-5505

[7]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 4171–4186. doi: 10.18653/v1/N19-1423

[8]. Tian, L., Moore, J., & Lai, C. (2016, December). Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 565-572). IEEE. doi: 10.1109/SLT.2016.7846319

[9]. Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., & Luo, J. (2020, July). A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3025-3035). doi:10.18653/v1/2020.acl-main.273

[10]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. doi:10.1145/3065386

[11]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

[12]. Ermatita, E., Sanmorino, A., Samsuryadi, S., & Rini, D. P. (2022). Analyzing Factors Contributing to Research Performance using Backpropagation Neural Network and Support Vector Machine. *KSII Transactions on Internet and Information Systems (TIIS)*, *16*(1), 153-172. doi:10.3837/tiis.2022.01.009

[13]. Gehring, J., Miao, Y., Metze, F., & Waibel, A. (2013, May). Extracting deep bottleneck features using stacked auto-encoders. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3377-3381). IEEE. doi: 10.1109/ICASSP.2013.6638284

[14]. Mukaihara, K., Sakti, S., & Nakamura, S. (2017, September). Recognizing emotionally coloured dialogue speech using speaker-adapted DNN-CNN bottleneck features. In *International Conference on Speech and Computer* (pp. 632-641). Springer, Cham. doi:10.1007/978-3-319-66429-3_63

[15]. Song, Y., McLoughLin, I., & Dai, L. (2015, June). Deep bottleneck feature for image classification. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (pp. 491-494). doi:10.1145/2671188.2749314

[16]. Hong, H., Choi, Y., Hahn, S., Park, S. K., & Park, B. J. (2014). Nomogram for sample size calculation on a straightforward basis for the kappa statistic. *Annals of Epidemiology*, *24*(9), 673-680. doi:10.1016/j.annepidem.2014.06.097

[17]. Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian. ACM Transactions on Asian Language Information Processing, 6(4), 1–33. doi:10.1145/1316457.1316459

[18]. Pramanik, S., & Hussain, A. (2019). Text normalization using memory augmented neural networks. *Speech Communication*, *109*, 15-23. doi:10.1016/j.specom.2019.02.003