International Journal of Applied Engineering & Technology

Transformer Model in Music Mood Classification

Thoyyibah. T [1]

Faculty of Information Technology, University of Pamulang, South Tangerang, Banten, Indonesia
Computer Science Department, BINUS Graduate, Program Doctor of Computer Science Bina Nusantara University Jakarta

dosen01116@unpam.ac.id

Edi Abdurachman[2], Yaya Heryadi [3], Amalia Zahra[4]

Computer Science Department, BINUS Graduate, Program Doctor of Computer Science Bina Nusantara University Jakarta, Indonesia,

edia@binus.ac.id[2], YayaHeryadi@binus.edu[3], amalia.zahra@binus.edu[4]

Date of Submission:, 11th March 2022, Revised : 22nd May 2022, Accepted: 2nd June 2022

***Abstract* - In recent years, artificial intelligence technology has developed rapidly and has become an inseparable part of life. Artificial intelligence has driven many activities in various sub-fields using machine learning. One of them is in the field of music. Every human being loves music. Music has become a part of human life. Music provides its own atmosphere to the listener which is called mood. There are so many elements of music that can be processed in machine learning, such as lyrics and audio. Lyrics and audio are saved into a music dataset. The author uses 1181 data on Indonesian music in the era of the 70s and 80s. The dataset only takes the refrain. The author uses the Transformer model where the lyrics management uses word embedding and audio processing uses the chromagram feature.**

***Keywords*— Indonesian Music dataset, , Mood, Transformer**

INTRODUCTION

Every human being loves music. Music knows no age, young or old, children and youth. All ages love music. Music is a subject in machine learning that can be studied further [1]. Music has many features. Music features include Spectograms, Cromagrams, and posters. Indonesia is a rich country besides being rich in natural resources, Indonesia is rich in culture [2] and music. Musical mood is the mood of music listeners caused by listening to the music. At this time the labeling (category) of mood on music is done manually by music experts after listening to the music. The rapidly increasing volume of music data from various genres makes labeling music categories inefficient. Therefore, a method that can be used to automatically label music moods into several categories is needed to support the production process and music analysis. One method of automatic music mood classification is using a machine learning model that is trained in a supervised manner with the input of a number of samples that have been labeled musical moods by collecting datasets manually. Indonesian music in the 70s and 80s is music that has many genres.

Among them dangdut, ballads, pop and struggle. The dataset used by the author is a dataset that is processed by himself by downloading songs from YouTube, converting to mp3 and cutting the song by taking only the chorus.

LITERATURE RIVIEW

Some topics that can be used for the use of transformer models in classifying moods in music, include:

A. Attributes to Music

The main musical features can be grouped into several categories as follows [3].

1. Music content is a music feature that is extracted from music audio signals. The main features of music content are rhythm (rhythm), timbre (tone color), melody (linear succession in music), harmony (a series of chords), loudness (loudness), and lyrics (song lyrics).
2. Music context is a music feature that is extracted directly from music audio. The main features of the music context are: semantic labels, performer's reputation, album cover artwork or posters, background artists, and music videos.
3. Music user context is a feature that represents the dynamic context of listeners/music connoisseurs. Music user context features consist of mood (such as feeling sad or happy), activities, social context, spatial temporal context, and other psychological conditions.
4. Music user property is a feature that represents dynamic characteristics of listeners/music connoisseurs such as music tastes, user opinions of players. The music user property features consist of: music preference, music training, music experience, demographics, opinions about musicians or performances, and popularity of musicians.

B. Classification of Mood in Music

Each music sample used as training data is represented by several musical features. The training model is then used to predict the mood category of the music sample that does not have a musical mood label [4]. The mood categories used in labeling music can be seen in Table 1. In general, musical moods can be grouped into positive and negative moods. Furthermore, the category of positive mood can be further divided into several sub-moods from the highest level (excited) to the lowest (calm). Likewise, the negative mood category can be further divided into several sub-moods from the highest level (annoying) to the lowest (sleepy) [5][6]

Table 1
Mood Grouping [5]

Category	*Positive*	*Negative*
High	*Excited, Happy,* *Pleased*	*Annnoying, Angry,* *Nervous*
Low	*Relaxed, Peaceful,* *calm*	*Sad, Bored, Sleepy*

Many researches on mood music have been carried out, for example research that uses mood as an access point for lyrics and audio [7]. Creating a music song mood dataset based on happy, angry, sad and relax categories. The creation of this dataset is based on Tags. last. Fm [8]. Another study explains that music can produce moods that can find suitable tracks in catalog and film productions [9]. Besides that, mood can also be used through many lyrics in OPM songs, where the method used is TF-IDF and the Key Graph keyword algorithm with an accuracy rate of 80%[10]. The implementation of mood in music is also processed with 1000 songs that have been annotated which in this study used multivariant regression reaching R to .70 and .50 [11]. The use of the mirex dataset is also carried out for detection of genre, mood and compuser by using audio features that are converted to mel spectrogram and CNN (Convolutional Neural Network) [12]. Besides that, mood is also studied as a person's concept in carrying out tasks such as driving, sleeping, romantic, confident, noisy, lively, noisy or others with various genres of jazz, comedy, gospel, sentiment, electronica, R & B Sensual, Blue, Blues, populist. Mood research on artists, genres was also carried out using dataset sources from AllMusicGuide.com, epinions.com and Last.fm [13].

C. Transformer Model

In essence, to solve sequential problems such as sentences, a transformer is needed which is an artificial neural network architecture [14]. Transformers are part of the Natural Language Process (NLP) in machine learning. The transformer is in charge of connecting the encoder and decoder to the text data [15]. input and output are interconnected through context vectors. Autoencoder is able to convert input into output. Besides being used in NLP transformers, it is also used in computer vision [16].

Several studies have used transformers, namely the Transformer Model to detect emotions in a social media conversation, for example on Facebook with the categories of happy, sad and angry [17]. Machine learning that uses a transformer model with 1000+ transformers to analyze health problems [18]. A transformer-based trial predicts the next 10 weeks with influenza case subjects [19]. The transformer model is also used in the detection of abusive language in Indonesian online news commentary. Text classification is done with the category of offensive, normal or non-ovensive [20]. In the health world, the transformer model is used for energy monitoring using the concept of algorithms in machine learning [21]. This study describes the use of the transformer method to analyze a person's stress [22]. Figure 1 describes the process of encoder and decoder through the input process and produces output. Transformer is also a mechanism that studies contextual relationships between words [23].

In Transformer there are two mechanisms, namely:

1. Encoder

The encoder is used to read all text input at once. The encoder consists of a stack of identical layers. Each layer has two sub-layers, namely self-attention layer and feed-forward neural network. With a self-attention layer, the encoder can help nodes not only focus on the word they are viewing but also to get the semantic context of the word. Each position in the encoder can handle all positions in the previous layer in the encoder.

2. Decoder

The decoder functions to produce a predictive output sequence. The decoder also consists of a stack of identifiable layers. Each layer consists of two sub-layers like those of the encoder, with an additional attention layer between the two layers to help the current node get to the key content that needs attention by performing multi-head attention on the output of the encoder. Similar to the encoder, the self-attention layer in the decoder makes each position in the decoder able to handle all the previous and current positions.

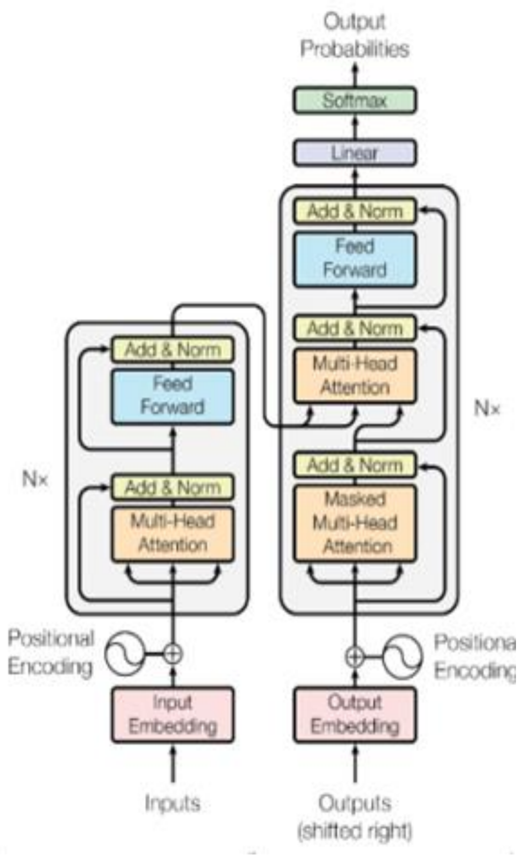


Figure 1. Transformer Model [23]

THE PROPOSED MODEL

The initial model or initial research framework used in the classification of moods with a database of Indonesian songs 70 and 80 eras is shown in Figure 2. The model in Figure 2 consists of the training music dataset and the testing music dataset. At the training stage, datasets are collected with data labels. Audio data will be processed into a vector value on the Chromagram, Lyrics data stored in excel will be processed using document embedding extraction. So that the resulting features in the form of chromagram features and lyric features. Chromagram data is the first training model. The lyric data is the second training model. The first and second training models will be combined into a trained ensemble model that will generate evaluation values. Likewise with the stages of data testing.

Figure 3 is a research framework with transformers. Where the stages are divided into 2, namely training music datasets and testing music datasets. This music dataset training involves music mood labels with audio and lyric data. Audio data will be processed into Chromagram. The lyric data will be processed through word embedding extraction. By using the transformer model, trained transformers are carried out which result in a performance evaluation.

Likewise with the music dataset testing stage which produces chromagram features and document embedding from audio and lyric data. Where the next step is a trained transformer model and generating a predicted Mood label.

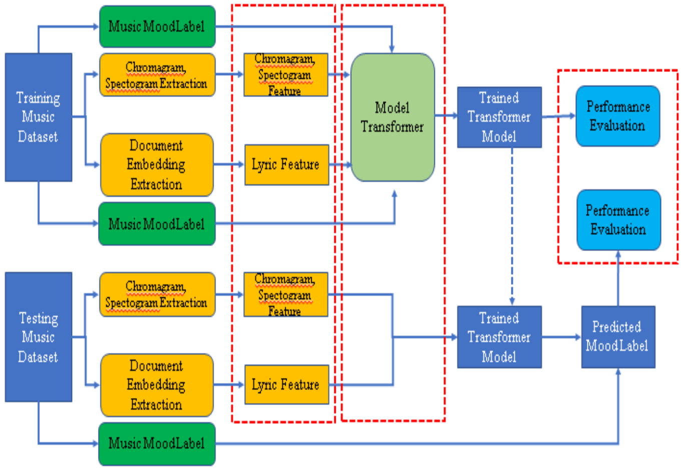

Music MoodLabel
Training Music Dataset
Chromagram, Spectogram Extraction
Chromagram, Spectogram Feature
Document Embedding Extraction
Lyric Feature
Music MoodLabel
Model Transformer
Trained Transformer Model
Performance Evaluation
Performance Evaluation
Testing Music Dataset
Chromagram, Spectogram Extraction
Chromagram, Spectogram Feature
Document Embedding Extraction
Lyric Feature
Trained Transformer Model
Predicted MoodLabel
Music MoodLabel

Figure 2. Initial research framework

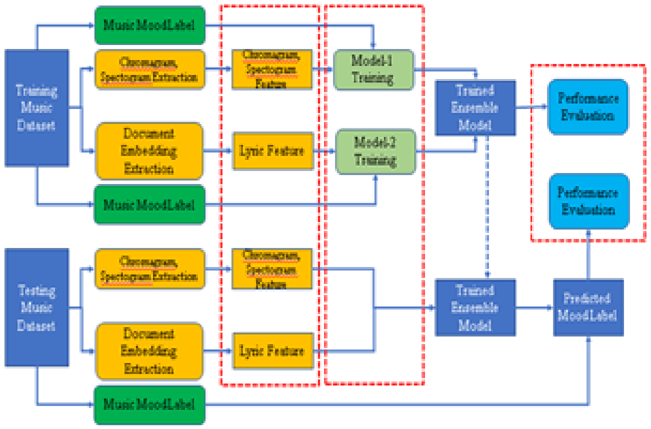


Figure 3. Research framework with transformers

RESULT AND DISCUSSION

Figure 4 is a Proposed Transformer Model Architecture as a Mood Classification. Where this image consists of input and output. Input has chromagram based feature vector and lyric based features. Figure 4 has many stages until it finally reaches the output. The first stage used is the same as the stages in the transformer, namely the vector encoder. The second stage is attention weight with weights of 0.5, 0.4, 0.3, 0.2, 0.1. The third stage is the context vector, the fourth stage is the decoder vector. Finally produces Output. The encoder and decoder get the input and convert it to a context vector and the decoder gets the context vector and converts it to an output.

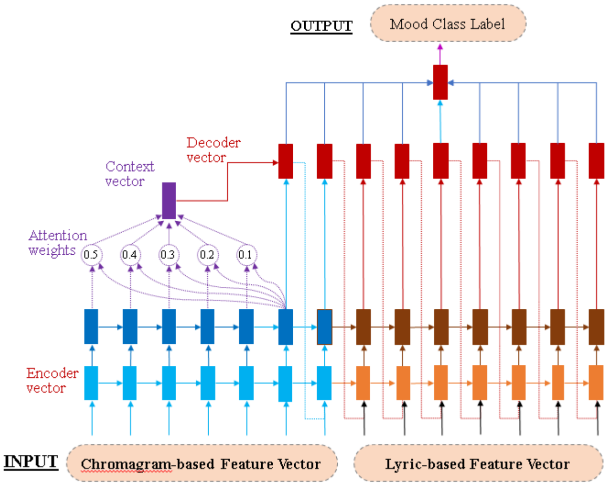


Figure 4. Proposed Transformer Model Architecture as Mood Classification

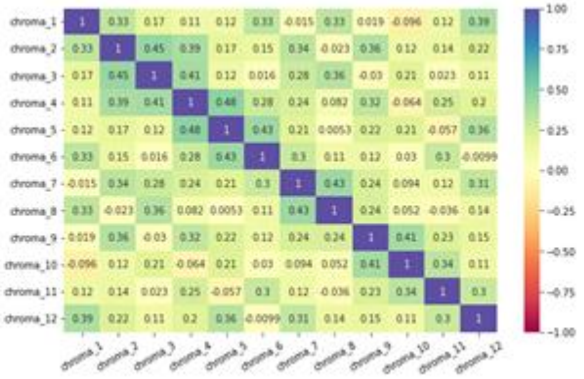
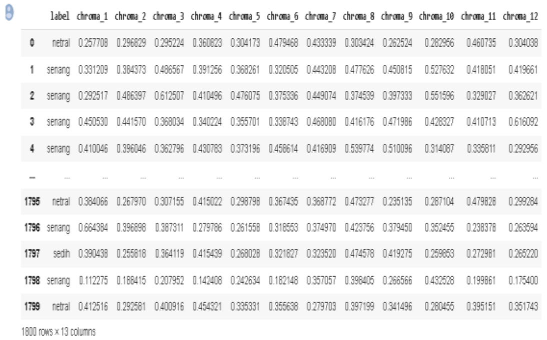


Figure 5. Correlation between chroma . columns

Figure 5 shows the degree of correlation between columns. Figure 5 only shows vector values in chroma 1 to chroma 12. Figure 6 is a data table of audio data containing vector values from the first chroma vector to the final data.



	label	chroma_1	chroma_2	chroma_3	chroma_4	chroma_5	chroma_6	chroma_7	chroma_8	chroma_9	chroma_10	chroma_11	chroma_12
0	netral	0.257708	0.296829	0.295224	0.360823	0.304173	0.479468	0.433339	0.303424	0.262524	0.282956	0.460735	0.304038
1	senang	0.331209	0.384373	0.486567	0.391256	0.368261	0.320505	0.443208	0.477626	0.450815	0.527632	0.418051	0.419661
2	senang	0.292517	0.486397	0.612507	0.410496	0.476075	0.375336	0.449074	0.374539	0.397333	0.551596	0.329027	0.362621
3	senang	0.450530	0.441570	0.368034	0.340224	0.355701	0.338743	0.468080	0.416176	0.471986	0.428327	0.410713	0.616092
4	senang	0.410046	0.396046	0.362796	0.430783	0.373196	0.458614	0.416909	0.539774	0.510096	0.314087	0.335811	0.292956
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1795	netral	0.384066	0.267970	0.307155	0.415022	0.298798	0.367435	0.368772	0.473277	0.235135	0.287104	0.479828	0.299284
1796	senang	0.664384	0.396898	0.387311	0.279786	0.261558	0.318553	0.374970	0.423756	0.379450	0.352455	0.238378	0.263594
1797	sedih	0.390438	0.255818	0.364119	0.415439	0.268028	0.321827	0.323520	0.474578	0.419275	0.259853	0.272981	0.265220
1798	senang	0.112275	0.188415	0.207952	0.142408	0.242634	0.182148	0.357057	0.398405	0.266566	0.432528	0.199861	0.175400
1799	netral	0.412516	0.292581	0.400916	0.454321	0.335331	0.355638	0.279703	0.397199	0.341496	0.280455	0.395151	0.351743

1800 rows × 13 columns

Figure 6. Data table of audio data

Figure 7 is a visualization of the length of the lyrics from the text data. To train a model, the size of the data is very important. Figure 7 is a visualization of the length of each lyric.

The longest size of the lyric data is around 40 words in the chorus and the smallest size is about 3 words. This needs to be processed again so that it gets the maximum value.

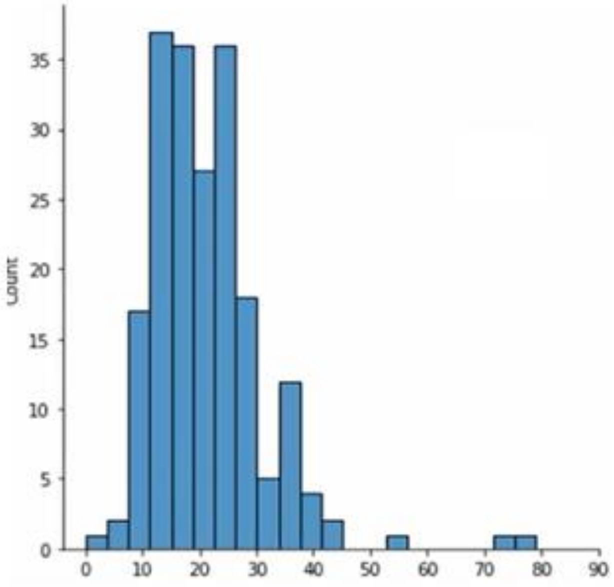


Figure 7. Visualization of text length

CONCLUSION

This study aims to develop a transformer model on music data using machine learning. Transformers are very important and need to be used in natural language processing with lyric data analysis on music datasets.

Transformer model used for processing voice data via audio. Lyrics data through chorus pieces stored in excel. For further research, you can use the intro or the ending section so that it can produce more datasets, high accuracy and more comparisons as well.

REFERENCES

[1] Alabi EO, Ogunajo FT, Fasae OD. Kalsifikasi Genre Musik Menggunakan Mesin dan Deep Learning Teknik: Tinjauan. Research Jet Journal of Analysis and INventions. Volume 3:2776-0960. 2022

[2] Anggeli P, Suroso, Agung MZ. Klasifikasi Alat Musik Tradisional dengan Metode Machine Learning dengan Librosa dan Tenserflow pada phyton. Jurnal sains komputer dan Informatika (J-SAKTI). Volume 5 No 2:2548-9771.2021

[3] Schedl, M., Gómez, E., & Urbano, J. Music Information Retrieval: Recent Developments and Applications. Foundations and Trends® in Information Retrieval (Vol. 8). .2014 .https://doi.org/10.1561/1500000042

[4] Panda RES. Emotion Based Analysis and Classification of Audio Music. Thesis.Institution granting the academic degree: University of Coimbra, Faculty of Sciences and Technology. 2019.

[5] R. E. Thayer. The Biopsychology of Mood and Arousal. New York, Ny: Oxford University Press. 1989.

[6] Bhattarai B and Lee J. Automatic Music Mood Detection Using Transfer Learning and Multilayer Perceptron. International Journal of Fuzzy Logic and Intelligent Systems 2019;19(2):88-96

[7] Hu X, Downie JS. Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. JCDL. 2010

[8] Cano E, Morisio M. Music Mood Dataset Creation Based Last. FM Tags. In Fourth International Conference on Artificial Intelligence and Applications (AIAP). Porto Institutional Repository. PP 15-16. DOI:10.5121/csit.2017. 70603. http://porto.polio.it/term_and_conditions.html

[9] Saari P, Barther M, Fazekas G, Eerola T, Sandler M. Semantic Models of Musical Mood: Comparison Between Crowd-Sourced And Curated Editorial Tags. 2013 IEEE International Conference on Multimedia & Expo (ICME2013) International Workshop on Affective Analysis in Multimedia (AAM),The Academy of Finland. 12033-76187.2013.

[10] V Emil L, Cabredo R. Lyric-Based Music Mood Recognition. Proceedings of the DLSU Research Congress Vol. 3. 2015

[11] Weninger F, Eyben F, Schuller. On-LIne Continous-Time Music Mood Regression with Deep Recurrent Neural Networks. The Euro pean Community's seventh FrameworkProgramme under grant aggrement no.338164 (ERC Starting Grant i HeARy). 2014

[12] Lidy T, Schindler. Paralel Convolutional Neural Network For Music Gennre and Mood Classification. Report number: Music Information Retrieval Evaluation eXchange (MIREX 2016). https://www.researchgate.net/publication/313895558_Parallel_Convolutional_Neural_Networks_for_Music_Genre_and_Mood_Classification

[13] Hu X and Downie JS. exploring mood metadata:Relationships with genre, artist and usage metadata http://ismir2007.ismir.net

[14] Luitse D and Denkena W. The Great Transforme: Examining The Role of Large Language Models in The Political Economy of AI. Big data and Society Journal. 2021

[15] Singla S and Ramachandra N. Comparative Analysis of Transformer Based Pre-Trained NLP Models. IJCSE International Journal of Compuetr Science and Engineering. Vol 8, Issue 11: 2347-2693. .2020.

[16] Ghojogh, Benyamin, and Ali Ghodsi. 2020. “Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey.” OSF Preprints. December 17. doi:10.31219/osf.io/m6gcn.

[17] Zhong P, Wang D, Mian C. Knowledge-Enriched Transformer for Emotion Detection In Textual Conversations. Empirical Methods in Natural Language Processing Proceedings. 2019. https://arxiv.org/abs/1909.10681

[18] Alqudsi A and El-Haq A. Application of Machine Learning in Transformer Health Index Prediction. Mdpi journal. 12, 2694. 2019. Doi:10.3390/en12142694.

[19] Wu N. Green B. Ben X. O'Banion S. Deep Transformer Models for Times Series Forecasting: The Influenza Prevalence Case. International Conference on Machine Learning. Vienna, Austria. 2020

[20] Rendragraha AD, BIjaksana MA, Romadhony A. Pendekatan metode transformers untuk deteksi bahasa kasar dalam komentar berita online Indonesia. E-Proceeding of Engineering : Vol. 8 No. 2. ISSN: 2355-9365.page 3385. 2021

[21] Tran QT, Davies K, Roose L.Machine learning for assessing the service transformes health using an energy monitor device. IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEEE). Volume 15. Issue 6. E-ISSN:2278-1676. P-ISSN: 2320-3331, Paper 01-06. www. Iosrjournals.org. 2020.

[22] Valencia F, Arcos H, Quilumba F. Comparison of Machine Learning Algorithms for the Prediction of Mechanical Stress in Three- Phase Power Transformer Winding Conductors. Hindawi Journal of Electrical and Computer Engineering. Article ID 4657696, Research Article 9 pages https://doi.org/10.1155/2021/4657696. 2021

[23] Vaswani, Ashish, Shazeer, Noam, Parmer, Niki, Uszkoreit, Jakob, Jones, Llion, Gomes, Aidan N, Kaiser, Lukasz and Polosukhin,lllia. Attention is all you need. In Advances in neural Information Processing Systems, PP 5998-6008. 2017

AUTHOR INFORMATION

Thoyyibah. T is a computer science doctoral student at Bina Nusantara University. He has presented several studies in international and national research conferences. Thoyyibah..T obtained a bachelor's degree in informatics engineering at the Syarif Hidayatullah State Islamic University, Jakarta and obtained a master's degree in computer science from the Bogor Agricultural Institute. In addition to his outstanding professional experience in teaching at Pamulang University, South Tangerang. Thoyyibah..T also often participates in various other campus activities, such as an independent campus and research for novice lecturers.

Edi Abdurachman is a professor in statistics inaugurated at Bina Nusantara University in 2009. He earned a Doctorate in statistics in 1986 and a Master of Science (M.Sc) in statistical surveying in 1983 from Iowa State University, Ames, USA. ; Master of Science (MS) degree. and a degree in Agricultural Engineer (Ir) (Cum laude) in 1978 from the Bogor Agricultural University (IPB). He has long experience as a statistical consultant at national and international levels. He is often invited as a speaker in International Conferences and various other seminars. For example, the International Seminar on Statistics and Data Utilization for Agricultural Policies in Myanmar in 2005 and Economics Modeling for the Agricultural Sector. The Case of Predicting some Agricultural Strategic Commodities; The Office of Agricultural Economics (OAE) and JICA ASEAD Project in Bangkok, Thailand. He has published a lot of research in the field of statistics. He is also a member of the American Statistical Association, International Association of Engineers (IAENG), IEEE, APTIKOM and an honorary member of the Statistical Honor MU SIMA RHO Society. He also received the Presidential Award for 10 years Satyalencana, 20 years Satyalencana, 30 years Satyalencana; and Teaching Awards; Best Lecturer Award

Yaya Heryadi holds a Bachelor's degree in Statistics and Computation from the Institut Pertanian Bogor, a Master of Science from Indiana University at Bloomington, USA, and a Doctorate in Computer Science from the Universitas Indonesia. During his career as a researcher, he took some courses at the University of Kentucky at Lexington, USA, and the sandwich-like program at Michigan State University at East Lansing, USA. Currently he is a lecturer and researcher at the Doctor of Computer Science (DCS) Department, Binus Graduate Program, Bina Nusantara University with research interests in Artificial Intelligence, Data Science, Machine Learning/Deep Learning, Natural Language Processing, and Computer Vision. The results of his works have been published in a number of international conference proceedings, international journals, books, and intellectual property rights.

Amalia Zahra is a lecturer at the Master of Information Technology, Bina Nusantara University, Indonesia. She received her bachelor’s degree in computer science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master’s degree. Her PhD was obtained from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014.

Her research interests cover various fields in speech technology, such as speech recognition, spoken language identification, speaker verification, speech emotion recognition, and so on. Additionally, she also has interest in Natural Language Processing (NLP), computational linguistics, machine learning, and artificial intelligence.