ICIC Express Letters
Volume **16**, Number **7**, July **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **787–794**

EMOTIONS CLASSIFICATION FROM CONVERSATION USING BERT

Andry Chowanda

Computer Science Department
School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
achowanda@binus.edu

Received June 2021; accepted September 2021

Abstract. *This research proposes the optimal model for emotions classification from social conversation using BERT, a pre-trained model from The Bidirectional Transformers for Language Understanding. The Friends dataset is used to fine-tune the models. The dataset consists of utterances of a Sitcom movie conversation labelled with six basic emotions plus Neutral. Two architectures (BERT and XLNet) and four pre-trained models (bert-base-uncased, bert-large-uncased, xlnet-base-cased, xlnet-large-cased) were implemented to model the emotions classification from social conversation. To deal with imbalanced class in the dataset, the loss function was modified with a weighted loss where the weight is calculated based on the number of labels (i.e., emotions on each class). The results demonstrate that the best model was achieved by the one fine-tuned with BERT architecture and bert-large-uncased pre-trained model (24 layers, 1024 hidden, 16 heads, with a total of 336M parameters). In addition, the model achieved the training F1-Score of 97.101%, validation F1-Score of 96.011%, and loss of 0.030.*
Keywords: Emotions recognition, Social conversation, Deep learning, BERT

1. **Introduction.** Social conversation is one of the methods we as social animals use to interact with each other. During the social conversation, a human generally exchanges social signals with their body. The first interlocutor conveys their social signals both verbally and non-verbally. The other interlocutors perceive the social signals, interpret them, and then return the social exchange. The social interaction loop will continue until one of the interlocutors breaks the interaction [1, 2]. Recognizing, interpreting and returning those social signals are paramount skill-sets to be implemented to an intelligent system (e.g., intelligent virtual agents or intelligent virtual humans [1, 2]) in order to build a system that is able to interact with humans naturally. However, interacting with humans is a cumbersome task for not socially aware computers. Several methods have been proposed to solve the problems. Some of the methods result in a model with a good understanding in the verbal context of a social conversation but still lacking in the social skills [3, 4, 5]. The others result in a model that has a limited understanding of social cues during a social conversation [1, 2]. Several tasks can be done in this area to build a socially aware system. First, the social cues or signals conveyed by the interlocutor can be captured, extracted and processed to interpret the social signals. One of the important signals that can be perceived from the interlocutor's social signals is emotions. Emotions provide flavours and complex meaning to social conversations. A word can have multiple meanings given the emotions induced by the word. For example, the word "OK" can mean that the interlocutor is either happy or angry. During the social conversation, emotions can be interpreted from verbal (speech, text) and non-verbal cues (speech, text, facial expressions, gestures). Machine and deep learning methods are generally implemented to model emotion recognition (classification) from social conversation based on

DOI: 10.24507/icicel.16.07.787

the emotions classification model. There are several emotions classification models in the field of psychology (e.g., discrete and dimensional model, see [6]) that can be used as the classification group. The most well-known emotions model used to model the emotions recognition from the social conversation is the discrete model, for example, Ekman's basic emotions [7].

This paper aims to build emotions recognition models from the social conversation (i.e., text modality) that can be used generally in any social conversation situation and are not limited to only a small set of social conversations and topics. This paper proposed several models trained with BERT (Bidirectional Encoder Representation from Transformers) architecture to model an emotions classifier from a social conversation with text modality. In addition, several pre-trained models were also implemented to accelerate the training speed and fine-tune the model with specific tasks proposed in this paper. The dataset used in this paper is the Friends dataset [8], where the dataset was collected from a Sitcom movie, Friends and annotated using Six Basic Emotions plus Neutral. The results demonstrate that the best model achieved by setting 2 (SET-2) with BERT architecture with the bert-large-uncased pre-trained model (24 layers, 1024 hidden, 16 heads, with a total of 336M parameters). The model achieved the training F1-Score of 97.101%, validation F1-Score of 96.011%, and loss of 0.030. The rest of this paper is organized as follows. The related work in recognizing emotions from the social conversation is thoroughly reviewed in the next section. Section "Emotions Classification from Conversation" comprehensively illustrates the methods proposed in this paper. The results then thoroughly presented and discussed in the "Results and Discussion" section. Finally, the results are concluded, and the future work direction is presented in the last section.

2. **Related Work.** The emotions recognition task has been an appealing topic to researchers during the past decades. Emotions from a human can be recognized from multiple sources of social signals conveyed by a human when interacting with each other. The social signals can be captured using sensors like a camera, microphone, and other biosensors [9, 10, 11]. The signals then are interpreted using appropriate algorithms or methods depending on the nature of the signals captured by the sensors [9, 10, 11, 12]. For example, signals captured by the camera (e.g., facial expressions) can be processed using algorithms or methods in the computer vision area, while signals captured by microphones (e.g., speech) can be processed using algorithms or methods that exist in the natural language processing (for the spoken words) and signal processing (for the prosody). Before deep learning methods flourished these days, features representation such as N-Gram, Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words was the standard representation for text features. The features were usually trained with support vector machine, logistic regression, or classification tree (e.g., XGBoost or random forest) to solve the classification problems in the text area. Nowadays, deep learned vector representations are the most popular features representations for text, trained with deep learning techniques [13, 14]. Deep learning architectures improved the model's accuracy compared to the one trained with conventional machine learning. Kratzwald et al. [15] argued that the models trained with deep learning algorithms (e.g., Long Short-Term Memory (LSTM) or Bi-LSTM) provide improvement of 23.2% in the F1-Score compared to the one trained with the conventional machine learning algorithms (e.g., Support Vector Machines (SVM) or random forests) in the classification tasks. Kratzwald et al. [15] achieved the best of F1-Score of 68.8% in the text classification model trained with a pre-trained model of Bi-LSTM.

Chowanda et al. [16] explored several conventional machine learning and deep learning algorithms to train models for emotions recognition from social media (i.e., text features) using the AffectiveTweets dataset. The best model was achieved by the one trained with the generalized linear model algorithm with the accuracy score of 92% and F1-Score of

90.1%. Su et al. [17] implemented two deep learning algorithms, namely LSTM and CNN, to model emotions recognition from text using the NLPCC-MHMC-TE database. The best model was trained with the LSTM architecture with 128 hidden nodes (Accuracy = 70.66%). In the second place, the model trained with Convolutional Neural Networks (CNN) with 100 features maps and kernel size of 2 achieved an accuracy score of 65.33%. Several deep learning architectures have some drawbacks in the training process. The recurrent neutral networks architectures such as LSTM and Bi-LSTM are superior in dealing with temporal information. However, the networks require a massive amount of computational power as the process cannot be paralleled. Moreover, due to the complexity of the networks, it can lead to a diminishing or exploding gradient problem. The state of the art of text classification tasks (including emotions recognition from text) is the pre-trained model of Transformer variations, such as Bidirectional Encoder Representations from Transformers (BERT) [18], Robustly Optimized BERT Pretraining Approach (RoBERTa) [19], and A Lite BERT for Self-supervised Learning of Language Representations (ALBERT) [20]. Huang et al. [21] fine-tuned emotions recognition model using EmoPush and Friends dataset using a pre-trained of BERT-Large. The model achieved an accuracy score of 88.5% and 86% for a model trained with the EmoPush dataset and Friends dataset. Using the same dataset and the modification of the BERT model (EmotionX-KU), Yang et al. [22] achieved the best F1-Score of 86.3% and 78.4% for a model trained with EmoPush dataset and Friends dataset, respectively. The researcher also implements the variation of BERT models, such as RoBERTa and ALBERT. Pant et al. [23] combined RoBERTa and ALBERT and achieved an accuracy of 85.55% and F1-Score of 55.8% for the classification task. Acheampong et al. [25] also implemented RoBERTa pre-trained model with steps of 500K and achieved 94.6%, 90.2%, and 96.4% for models trained with the SQuAD (conversation) dataset, MNLI-m dataset, and SST-2 dataset, respectively. With the combination of BERT models, this research aims to build emotions recognition models from the social conversation (i.e., text modality) that can be used generally in any social conversation situation and are not limited to only a small set of social conversations and topics.

3. **Emotions Classification from Conversation.** The goal of this research is to build an optimal emotions classification model from the conversation by fine-tuning pre-trained models from BERT [18]. The models were trained using labelled social conversation from Friends dataset [8]. The dataset consists of utterances from a Sitcom movie, Friends. The utterances were labelled with Six Basic Emotions (i.e., Anger, Disgust, Fear, Joy, Sadness, Surprise) plus Neutral. Table 1 illustrates the settings and profile of the dataset used in this research. In total, 9,508 labelled utterances are used in the training process, where 8,544 (±90%) are used in the training phase and 964 (±10%) are used in the validation phase. As a result, the dataset is highly imbalanced in the number of utterances among the classes. For example, Neutral Class has the most significant number of utterances (5,243),

Table 1. Dataset overview

No	Emotions	All	Train	Validation
1	Neutral	5,243	4,752	491
2	Anger	598	513	85
3	Disgust	263	240	23
4	Fear	214	185	29
5	Joy	1,406	1,283	123
6	Sadness	413	351	62
7	Surprise	1,371	1,220	151
Total		**9,508**	**8,544**	**964**

and Fear Class has the lowest number of utterances (214). Hence, the weighted loss function was implemented in this research (see Equation (2)) to deal with the imbalanced dataset. The approach was chosen as the alternative to deal with the imbalanced dataset as sampling techniques did not provide significant improvements to the model.

A pre-trained model from the Bidirectional Encoder Representation from Transformers (BERT) [18] was implemented in research to build the emotions recognition model. BERT is a bidirectional training of Transformer to model languages. The model trained with BERT demonstrates to have a better and deeper sense of language context, compared to the other models [18]. Figure 1 illustrates the general model of BERT. The words (e.g., W_1-W_n) are converted into a vector form, called embedding. The attention mechanism from the Encoder layers of Transformer then learns the contextual relations of the words within the text. The Encoder layers read the words bidirectionally from the sentences. Finally, to support the classification task, a classification layer is added to the architecture. The classification layer provides the probability of the output based on the input processed by using the Softmax function. Equation (1) shows the probability of the output by calculating the Sotfmax of the word representation C multiplied by the weight transformed W^T plus the bias b.

$$P = softmax\left(CW^T + b\right) \tag{1}$$

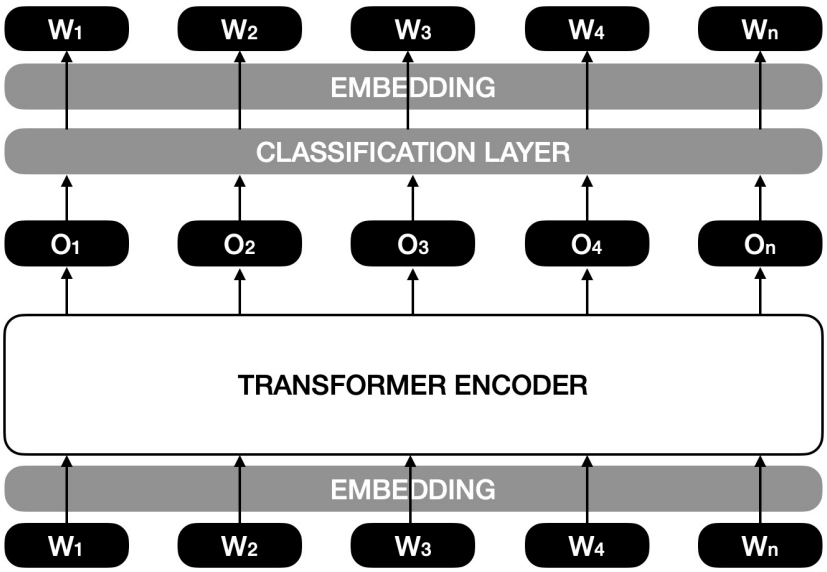

FIGURE 1. BERT model [23]

Four settings were proposed in this research using pre-trained models. Table 2 demonstrates the training architecture, pre-trained model, best learning rate, and best batch in each setting. Initially, several training processes were performed using several learning rates and batches settings to find the optimal hyper-parameters for each architecture. Two architectures and four pre-trained models (two models per architecture) were used to train the model. Original BERT architecture [18], and the autoregressive model, XL-Net architecture [24] were used in this research to fine-tune the pre-trained model of bert-base-uncased (SET-1), bert-large-uncased (SET-2), xlnet-base-cased (SET-3), and xlnet-large-cased (SET-4). The bert-base-uncased pre-trained model was trained in 12 layers, 768 hidden, 12 heads of Transformer architecture with 110M parameters. The bert-large-uncased pre-trained model was trained with larger architecture compared to the bert-base-uncased (24 layers, 1024 hidden, 16 heads, with a total of 336M parameters). The xlnet-base-cased pre-trained model was trained in 12 layers, 768 hidden, 12 heads of XLNet architecture with 110M parameters. In contrast, the xlnet-large-cased pre-trained model trained 24 layers, 1024 hidden, 16 heads of XLNet architecture with

Table 2. The best training hyper-parameters settings

No	Setting	Architecture	Pretrained model	Learning rate	Batch
1	SET-1	BERT	bert-base-uncased	1e-4	48
2	SET-2		bert-large-uncased	2e-5	48
3	SET-3	XLNet	xlnet-base-cased	2e-5	32
4	SET-4		xlnet-large-cased	5e-5	16

340M parameters. The weighted loss was implemented in this research to deal with the imbalanced class problem in the dataset. Equation (2) illustrates the formula for the weighted loss L. The loss is weighted based on the number of corresponding emotions for each class (w_c). The weight w_c is calculated based on the number of labels (i.e., emotions on each class).

$$L = -\frac{1}{\sum_{i=1}^{N} {w_c}^{(i)}} \sum_{i=1}^{N} \log\left(w_c \hat{P}_c^{(i)}\right) \tag{2}$$

4. **Results and Discussion.** Two architectures and four pre-trained models with the best learning rate and batch value are applied to training the emotions models from social conversation. All the training were limited to 20 epochs, as the models trained with pre-trained models. Table 3 illustrates the best results achieved for each setting. Overall, the results are quite similar among the settings. The best F1-Score in validation (96.241%) and loss (0.023) is achieved by the one trained with BERT architecture using the bert-base-uncased pre-trained model, with a learning rate of 1e-4 and batch of 48 (SET-1). However, the best F1-Score in training (97.101%) was achieved by a model trained with BERT architecture using the bert-large-uncased pre-trained model, with a learning rate of 2e-5 and batch of 48 (SET-2). Figure 2 demonstrates the training and validation results of all architectures fine-tuned with pre-trained models. The upper left of the figure shows the training and validation F1-Score of the model trained with SET-1 (bert-base-uncased). The best model resulted in 96.851%, 96.241%, and 0.023 for training F1-Score, validation F1-Score, and loss, respectively. The worst model resulted in 70.179%, 62.958%, and 0.237 for training F1-Score, validation F1-Score, and loss, respectively. The upper right of Figure 2 illustrates the training and validation F1-Score of the model trained with SET-2 (bert-large-uncased). The best model resulted in 97.101%, 96.011%, and 0.030 for training F1-Score, validation F1-Score, and loss, respectively. The worst model resulted in 69.023%, 61.415%, and 0.224 for training F1-Score, validation F1-Score, and loss, respectively.

Table 3. Best results

No	Model	F1-train	F1-validation	Loss
1	SET-1 (BERT-BASE)	96.851%	**96.241%**	**0.023**
2	SET-2 (BERT-LARGE)	**97.101%**	96.011%	0.030
3	SET-3 (XLNET-BASE)	96.172%	95.389%	0.034
4	SET-4 (XLNET-LARGE)	96.196%	95.677%	0.032

The results of the models trained with XLNet architecture are shown in the bottom part of Figure 2. The lower left of the figure demonstrates the training and validation F1-Score of the model trained with SET-3 (xlnet-base-cased). The best model resulted in 96.172%, 95.389%, and 0.034 for training F1-Score, validation F1-Score, and loss, respectively. The worst model resulted in 67.844%, 61.093%, and 0.235 for training F1-Score, validation F1-Score, and loss, respectively. The lower right of the figure shows the training and validation F1-Score of the model trained with SET-4 (xlnet-large-cased). The best model

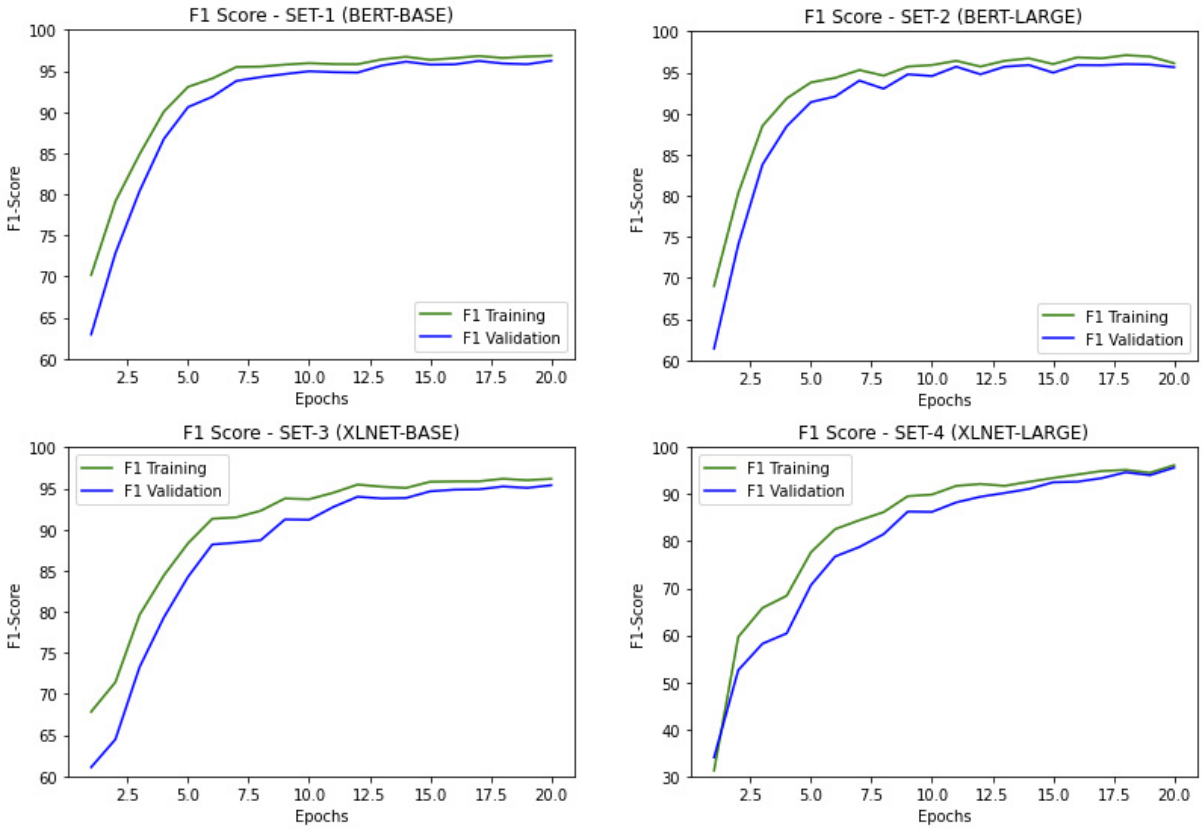

FIGURE 2. Training and validation F1-Score

TABLE 4. Prediction results samples

Utterance	True	Predicted			
		SET-1	SET-2	SET-3	SET-4
why do all youre coffee mugs have numbers on the bottom?	Surprise	Surprise	Surprise	Neutral	Surprise
okay, so what you used to have with rachel, is what ive got with alice	Joy	Neutral	Joy	Neutral	Joy
its out there man! ive seen it! i got it!!	Joy	Joy	Joy	Joy	Joy
all right, man!!	Joy	Joy	Joy	Joy	Joy
okay, y' know what? there is no more left, left!	Anger	Anger	Anger	Anger	Anger
oh my god! i overslept! i was supposed to be on the set a half an hour ago! i gotta get out of here!	Fear	Surprise	Surprise	Surprise	Surprise
oh, its bad. its really bad. the only thing in there that isnt burned is an ass. which i do	Sadness	Disgust	Sadness	Sadness	Sadness

resulted in 96.196%, 95.677%, and 0.032 for training F1-Score, validation F1-Score, and loss, respectively. The worst model resulted in 31.194%, 34.012%, and 0.701 for training F1-Score, validation F1-Score, and loss, respectively. Table 4 demonstrates the examples of prediction results from all the models. The model trained with setting SET-2 and SET-4 provides the best results in predicting the emotions from given utterances. In summary,

the best model trained with the BERT architecture, fine-tuned with bert-large-uncased with F1-Score of training of 97.101%. Moreover, SET-2 provides a more accurate prediction given a sample of utterances. In this research, training using BERT models (i.e., BASE and LARGE) provides the best results than the models that implement autoregressive language models (i.e., XLNet). Moreover, the weighted loss improved the training and validation results, albeit the imbalanced problem in the dataset.

5. **Conclusion and Future Work.** This research proposes the optimal model for emotions recognition from a social conversation with Transformer and XLNet architecture using the Friends dataset. The dataset provides labelled utterances from a Sitcom movie, Friends. Two architectures (XLNet and BERT) and four pre-trained models (bert-base-uncased, bert-large-uncased, xlnet-base-cased, and xlnet-large-cased) were used to fine-tune the emotions recognition model. Due to the nature of the dataset where the classes are highly imbalanced, a weighted loss formula implemented during the fine-tuning process. Overall, the best model is achieved by the one trained with setting SET-2 with BERT architecture with the bert-large-uncased pre-trained model (24 layers, 1024 hidden, 16 heads, with a total of 336M parameters). The model achieved the training F1-Score of 97.101%, validation F1-Score of 96.011%, and loss of 0.030. For future direction research, more datasets can be used to enrich the current dataset. Datasets using local language (e.g., Indonesian language) also can be used to model emotions recognition from a social conversation in the Indonesian language. Moreover, more architectures and pre-trained models also can be explored with different combinations of hyper-parameters to find more optimal results. At the feature level, multimodal features (e.g., text, speech, and image) can be implemented to enrich the information gathered from the social conversation. Moreover, temporal information also can be extracted from the current features to provide more accurate and rich information from the social conversation.

Acknowledgment. We also would like to extend our gratitude to NVIDIA Corporation with the donation of the Titan V Pascal GPU used for this research.

REFERENCES

[1] A. Chowanda, P. Blanchfield, M. Flintham and M. Valstar, ERiSA: Building emotionally realistic social game-agents companions, *International Conference on Intelligent Virtual Agents*, pp.134-143, 2014.

[2] A. Chowanda, M. Flintham, P. Blanchfield and M. Valstar, Playing with social and emotional game companions, *International Conference on Intelligent Virtual Agents*, pp.85-95, 2016.

[3] W. Zhu, A. Chowanda and M. Valstar, Topic switch models for dialogue management in virtual humans, *International Conference on Intelligent Virtual Agents*, pp.407-411, 2016.

[4] R. Sutoyo, A. Chowanda, A. Kurniati and R. Wongso, Designing an emotionally realistic chatbot framework to enhance its believability with aiml and information states, *Procedia Computer Science*, vol.157, pp.621-628, 2019.

[5] A. Chowanda and A. D. Chowanda, Recurrent neural network to deep learn conversation in Indonesian, *Procedia Computer Science*, vol.116, pp.579-586, 2017.

[6] E. Harmon-Jones, C. Harmon-Jones and E. Summerell, On the importance of both dimensional and discrete models of emotion, *Behavioral Sciences*, vol.7, no.4, DOI: 10.3390/bs7040066, 2017.

[7] P. Ekman, Basic emotions, in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power (eds.), John Wiley & Sons, Ltd., DOI: 10.1002/0470013494.ch3, 1999.

[8] B. Shmueli and L.-W. Ku, SocialNLP EmotionX 2019 challenge overview: Predicting emotions in spoken dialogues and chats, *arXiv Preprint*, arXiv: 1909.07734, 2019.

[9] A. Vinciarelli, M. Pantic and H. Bourlard, Social signal processing: Survey of an emerging domain, *Image and Vision Computing*, vol.27, no.12, pp.1743-1759, 2009.

[10] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico and M. Schroeder, Bridging the gap between social animal and unsocial machine: A survey of social signal processing, *IEEE Trans. Affective Computing*, vol.3, no.1, pp.69-87, 2011.

[11] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic and A. Vinciarelli, *Social Signal Processing*, Cambridge University Press, 2017.

[12] Y. Omae, M. Mori, T. Akiduki and H. Takahashi, A novel deep learning optimization algorithm for human motions anomaly detection, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.199-208, 2019.
[13] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text, *Journal of Artificial Intelligence Research*, vol.30, pp.457-500, 2007.
[14] S. Poria, N. Majumder, R. Mihalcea and E. Hovy, Emotion recognition in conversation: Research challenges, datasets, and recent advances, *IEEE Access*, vol.7, pp.100943-100953, 2019.
[15] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel and H. Prendinger, Deep learning for affective computing: Text-based emotion recognition in decision support, *Decision Support Systems*, vol.115, pp.24-35, 2018.
[16] A. Chowanda, R. Sutoyo, S. Tanachutiwat et al., Exploring text-based emotions recognition machine learning techniques on social media conversation, *Procedia Computer Science*, vol.179, pp.821-828, 2021.
[17] M.-H. Su, C.-H. Wu, K.-Y. Huang and Q.-B. Hong, LSTM-based text emotion recognition using semantic and emotional word vectors, *2018 1st Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp.1-6, 2018.
[18] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of NAACL-HLT*, pp.4171-4186, 2019.
[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, *arXiv Preprint*, arXiv: 1907.11692, 2019.
[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, *The International Conference on Learning Representations (ICLR)*, 2020.
[21] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu and Y.-S. Chen, EmotionX-IDEA: Emotion BERT – An affectional model for conversation, *arXiv Preprint*, arXiv: 1908.06264, 2019.
[22] K. Yang, D. Lee, T. Whang, S. Lee and H. Lim, EmotionX-KU: BERT-max based contextual emotion classifier, *arXiv Preprint*, arXiv: 1906.11565, 2019.
[23] K. Pant, T. Dadu and R. Mamidi, BERT-based ensembles for modeling disclosure and support in conversational social media text, *AffCon@AAAI*, pp.130-139, 2020.
[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems*, vol.32, 2019.
[25] F. A. Acheampong, H. Nunoo-Mensah and W. Chen, Transformer models for text-based emotion detection: A review of BERT-based approaches, *Artificial Intelligence Review*, vol.54, no.8, pp.5789-5829, 2021.