DistilBERT with Adam Optimizer Tuning for Text-based Emotion Detection

Farica Perdana Putri[1]
[1]Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
farica@umn.ac.id

Accepted 31 May 2023
Approved 11 July 2023

Abstract— **Emotion detection (ED) refers to identifying individual emotions or feelings, such as happiness, sadness, disappointment, fear, etc. The classic machine learning technique still relies on feature engineering, which makes it difficult to convey the meaning of words. Deep learning-based algorithms have recently been shown to be beneficial for emotion detection because they require only a simple feature creation process. Transfer learning is an approach that uses data similarities, data distribution, models, tasks, and other factors to apply knowledge learned in one domain to a new domain. This study is to shed light on the fine-tuned models' efficacy in detecting emotions from the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset. In order to optimize the model, we conducted the hyperparameters tuning on the Adam optimizer in DistilBERT. The experiment examined the moment estimators and learning rate of the Adam optimizer. The effect of the parameters on training and validation accuracy was presented and analyzed. Adam optimization first-moment estimators provide more robust convergence to the model during the training process as their value approaches one. The testing results of emotion detection is 97.14%..**

Keywords— ***emotion detection; transfer learning; Adam optimizer; DistilBERT.***

I. Introduction

The advancement of internet technology results in a rising amount of multimedia data. Multimedia data includes text, speech, images, and video and is produced and contains a vast amount of information. One type of digital content that is shared online is textual data. Social media generates millions of messages per day, necessitating immediate unstructured data processing [1]–[3]. Opinion mining has become essential in recent studies as a result of this occurrence [4]. Users can express or share their actions, thoughts, views, and emotions on social media [5]. As a result, a lot of businesses use this opportunity to analyze user-generated data as a source of analysis for internal decision-making [6], [7].

Emotion detection (ED) refers to identifying individual emotions or feelings, such as happiness, sadness, disappointment, fear, etc. Emotions can be detected in a person's voice, facial expressions, movements, and writing. This is based on the opinion that when someone is sad, he will use words that are not encouraging. Emotional acceptance and detection are used in various fields, such as healthcare, education, and advertising [8]. In healthcare, emotions or feelings can affect a person's health. Depressed emotions or depression can drain mental energy. Therefore, it harms the body and causes a decrease in health. Meanwhile, a joyful atmosphere can help restore one's health. In education, emotions play an essential role in the quality and quantity of student learning. Brain activity will increase when children experience positive emotions so that they can concentrate better. Emotional marketing is currently in great demand by companies because of its effectiveness in attracting consumer interest. Emotional marketing is a marketing and advertising effort that uses emotions to make audiences pay more attention to, remember, and make purchases. Many researchers are interested in improving the relationship between humans and computers using emotional extraction on social media, such as Twitter, Instagram, YouTube, Facebook, etc. [9]. Text-based emotion detection can be applied using NLP techniques to analyze the semantic meaning of a text.

Different from emotion detection research from faces [8] and voices [10], text detection is more challenging because it requires careful modeling of text since words associate with different emotions in different contexts with varying levels of magnitude making the identification of words for document representation more challenging [11]. Online social media content is short, informal, and unstructured text comprising incomplete and misspelled words, abbreviations, acronyms, and special characters. Moreover, like most NLP, we face problems such as negative words and anaphora which can change the meaning of sentences. An example of a negative word:

'I am not sad'. Sad is a word that shows implicit emotion, but the presence of the word not before sad makes the meaning of the sentence change to be similar to happy. The classification could be wrong if the computer cannot grasp this meaning. Anaphora refers to using a grammatical substitute like a pro-noun or pro-verb to denote a preceding word or a group of words.

Previous studies recognized human emotion from text using a rule-based and feature representation technique. The rules of emotion are then derived utilizing statistics, linguistics, and computation techniques. The best rules are chosen later. The rules are then applied to emotion datasets to determine emotion labels. Dibyendu et al. [12] proposed semantic rules to identify emotion at the sentence level. The method examines emotion words and phrasal verbs, as well as negation words, and performs better than previous approaches. presents a rule-based technique for detecting the emotion or mood of a tweet and categorizing it in the proper emotional category. The drawbacks of rule-based systems are a lack of contextual meaning and a lexicon with insufficient words.

The feature-based method relies heavily on machine learning techniques. It has improved classification accuracy by using appropriate models and characteristics to categorize text, but its efficiency is often lower than that based on the emotion dictionary [5]. The classic machine learning technique still relies on feature engineering [13], which makes it difficult to convey the meaning of words. Singh et al. [14] proposed the feature combination using semantic and statistical of words during the selection of significant features to construct the word vector. The framework consists of two stages: semantic-based to extract the meaningful words using POS tagger and statistical-based to remove the weak semantic feature using the Chi-square method.

Pang et al. [15] construct the topic-level feature space by grouping semantically related words. They develop the weighted labeled topic model (WLTM) and X-term emotion-topic model (XETM) to detect emotions toward certain topics. WLTM defines one-to-many mapping between emotions and multiple topics. XETM uses emotion distributions of labeled documents to constrain the topic probability of each feature during the training process. J.Guo [16] utilized deep learning and natural language processing to detect humans' emotions in text. They combined the questionnaire-based approach and the text-analysis-based approach features as feature vectors in the prior classifier. The method produced a human detection emotion rate of 97.22% and a classification emotion rate of 98.02%. Anzum and Gavrilova [4] introduced a novel approach of feature representation from Twitter based on Genetic Algorithm (GA). It is composed of stylistic, sentiment, and linguistic features extracted from tweets data. They used an ensemble classifier with weights optimized by GA to increase the detection accuracy.

Deep learning-based algorithms have recently been shown to be beneficial for emotion detection because they require only a simple feature creation process. One of the primary benefits of deep learning-based text mining approaches is efficient feature engineering. Transfer learning is an approach that uses data similarities, data distribution, models, tasks, and other factors to apply knowledge learnt in one domain to a new domain [17]. The labeled data can be utilized to construct the model and employed in the target domain data to enhance the annotation of the target data via transfer learning. Transfer learning starts with a pre-trained model, and fine-tuning involves further training the pre-trained model on the new task by updating its weights. Thus, the purpose of this study is to shed light on the fine-tuned models' efficacy in detecting emotions from the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [18].

II. Method

1. *Data Acquisition*

The ISEAR project was coordinated by Klaus R. Scherer and Harald Wallbott during the 1990s. Students, both psychologists and non-psychologists, were asked to describe scenarios in which they had felt all seven major emotions (joy, fear, anger, sadness, disgust, shame, and guilt). Thus, the final data set included reports on seven emotions from nearly 3000 respondents from 37 countries across all five continents. As we can see in Figure 1, the distribution of classes in the dataset is balanced.

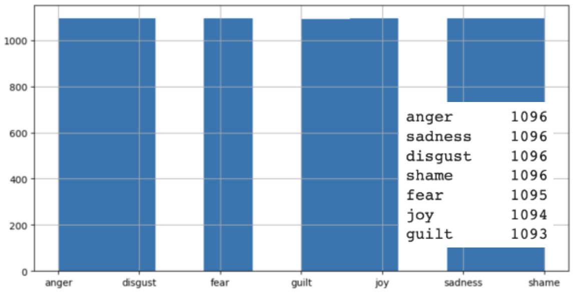


Fig. 1. Distribution of classes in the ISEAR dataset

2. *Data Preprocessing*

In this preprocessing step, the null values are checked and encode the classes from categorical to numerical values used in the fine-tuning process. The training and testing data is 80% and 20%, respectively. We used 20% of training data as validation data. The details of the number of training and testing data each class are described in Table 1.

TABLE I. NUMBER OF TRAINING AND TESTING DATA PER CLASS

	Number of training data	6132	Number of testing data	1534
Classes	Joy	875	Joy	219
	Fear	876	Fear	219
	Anger	877	Anger	219
	Sadness	877	Sadness	219
	Disgust	877	Disgust	219
	Shame	876	Shame	220
	Guilt	874	Guilt	219

3. *Fine-tuning Model*

In this work, human emotion detection is developed using fine-tuned DistilBERT. BERT [19] is a bidirectional transformer that was trained on a huge corpus of the Toronto Book Corpus and Wikipedia using a combination of masked language modeling objectives and next sentence prediction. BERT, in contrast to current language representation models, is intended to pre-train deep bidirectional representations from unlabeled text by conditioning on both left and right context in all layers.

DistillBERT is a lightweight, simple, inexpensive transformer model based on the BERT architecture. DistillBERT reduces 40% of the size of the BERT during the training phase using the knowledge distillation technique while 60% faster at inference time. Knowledge distillation is a technique used to transfer knowledge from a larger model, called the teacher, to a smaller model, called the student. DistilBERT retains 97% of the language understanding capabilities [20].

4. *Adam Optimizer*

In order to optimize the model, we conducted the hyperparameters tuning on the Adam optimizer. Adam [21] is an adaptive learning rate optimization algorithm that’s been designed specifically for training deep neural networks. However, the hyperparameters have intuitive interpretations and typically require little tuning. The Adam update rule is as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t \qquad (1)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2 \qquad (2)$$

where *m* and *v* are moving averages, *g* is gradient on the current batch, *t* is the number of iterations and β_1 and β_2 are hyper-parameters of the algorithm. β_1 is the exponential decay rate of the first momentum estimate and β_2 is the exponential decay rate of the second momentum estimate. This value should be set close to 1.0 on problems with a sparse gradient (e.g. NLP and computer vision problems).

Moving averages are set to zero at the beginning of each iteration. By evaluating the predicted values of moving averages, one can deduce moving averages' correlation with moments. As a result, the correction step removes the bias from the first and second moments caused by the optimizer's bias toward zero due to our zero initialization. Following estimator bias adjustment, the expected value becomes the desired value. Eqs. (3) and (4) provide bias-corrected estimators for the first and second moments.

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad (3)$$

$$\widehat{v}_t = \frac{v_t}{1-\beta_2^t} \qquad (4)$$

The algorithm's last step is to employ moving averages to scale the learning rate independently for each parameter. To apply this step, compute the weight update using Eq. (5).

$$w_t = w_{t-1} - \alpha \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} \qquad (5)$$

where *w* is model weight, α is the learning rate or step size, *t* is the number of iterations, and ϵ is to prevent division by zero and the default value is 10^{-8}.

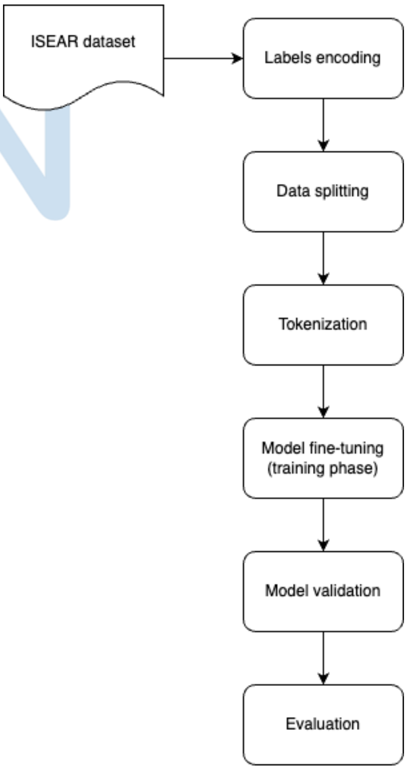
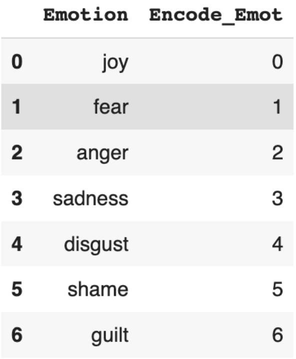


Fig. 2. Pipeline of the proposed system

III. Result and Discussion

The overall stages are explained for the proposed emotion detection system in the diagram in Figure 2. Firstly, the labels of the ISEAR dataset are transformed into numerical values ranges from 0 to 6 as depicted in Figure 3. There are seven classes include: joy, fear, sadness, shame, disgust, guilt and anger. Total number of samples is 7666 and divided into training, validation, and testing data based on the defined percentage.

Words in the sentence are tokenized using pre-trained uncased DistilBERT tokenizer specific for emotion detection. Tokenization is used in natural language processing to split paragraphs and sentences into smaller units that can be more easily assigned meaning. The results then forwarded to the model to perform the fine-tuning process. We created the customized model, by adding a drop out and a dense layer on top of DistilBERT to get the final output for the model.



	Emotion	Encode_Emot
0	joy	0
1	fear	1
2	anger	2
3	sadness	3
4	disgust	4
5	shame	5
6	guilt	6

Fig. 3. Encoded labels of the ISEAR dataset

During the validation stage, we pass the unseen data from the validation dataset to the model. This step determines how well the model performs on the unseen data before we evaluate it. During the validation stage, the weights of the model are not updated. Only the final output is compared to the actual value. This comparison is then used to calculate the accuracy of the model. Finally, the testing data is predicted using the best parameters.

The model parameters are described in Table 2. The number of batch sizes and epochs were determined by training the model with a learning rate of 0.00001 and default values of β_1 and β_2, which are 0.9 and 0.999, respectively. Different batch sizes (4, 8, and 16) and epochs (5, 10, and 15) have been tested. The results show that the best training and validation accuracy obtained from batch size is equal to 4 and epoch of 10, as shown in Table 3. Since the first and second moment estimators of the Adam optimizer can be evaluated for different learning rates, the objective is to compare the combination of different hyper-parameters in each training process. The accuracy of the training and validation processes for each possible hyperparameter is shown in Table 4.

Different learning rate parameters are taken into account while modifying the first and second moment estimations, which are compared among themselves. Option 1 is the best hyperparameters selection compared to other hyperparameters pairs. The overall training and validation accuracy become smaller while the learning rate is increased. A low learning rate would result in slower model training, requiring many parameter updates to reach the point of minimum. A high learning rate, on the other hand, would imply huge steps or abrupt modifications to the parameters, which frequently leads to divergence rather than convergence. Selected hyperparameters in Options 9-12, with a learning rate of 0.001, are not suitable for the network. It means the model get stuck in local optima because the learning parameter is too high. In most cases, the training accuracy is directly proportional to validation accuracy when β_1 is closer to 1. The increase in β_1 reduces the effect of gradients on the moving average of variance and provides more stable convergence. The testing accuracy using the best hyperparameters selection is 97.14%.

TABLE II. The Proposed DistilBERT Parameters

Parameter	Value
Batch size	[4, 8, 16]
Epochs	[5, 10, 15]
Tokenizer	Distilbert-base-uncased-emotion
Dropout	0.3
Optimizer	Adam
Loss function	Cross-entropy

TABLE III. The Results for Each Batch Size and Epoch Selection

Epoch	Batch Size	Training Accuracy	Validation Accuracy
5	4	93.65	98.68
5	8	93.37	98.84
5	16	85.57	88.68
10	**4**	**97.80**	**97.61**
10	8	97.73	97.68
10	16	97.39	97.35
15	4	98.10	96.60
15	8	98.03	96.77
15	16	97.86	95.60

TABLE IV. THE RESULTS FOR EACH HYPER-PARAMETER SELECTION

Option	Learning Rate	β_1	β_2	Training Accuracy	Validation Accuracy
1	**1e-05**	**0**	**0.99**	**98.57**	**96.39**
2	1e-05	0	0.999	97.57	95.33
3	1e-05	0.9	0.99	98.51	94.76
4	1e-05	0.9	0.999	97.79	97.21
5	1e-04	0	0.99	94.49	87.67
6	1e-04	0	0.999	92.03	78.37
7	1e-04	0.9	0.99	96.82	83.43
8	1e-04	0.9	0.999	95.48	80.90
9	1e-03	0	0.99	13.73	13.70
10	1e-03	0	0.999	13.98	13.87
11	1e-03	0.9	0.99	13.90	13.88
12	1e-03	0.9	0.999	14.14	13.70

.

IV. CONCLUSION

In this study, various hyperparameters of the Adam optimizer are tested on the ISEAR dataset to detect human emotion from text. According to the conducted experiments, the best hyperparameter selection was Option 1. We conclude that training and validation accuracy are lower as the learning rate increases. Adam optimization first-moment estimators provide more stable convergence to the model during the training process as their value is close to 1. For the future works.

REFERENCES

[1] A. R. Murthy and K. M. Anil Kumar, "A Review of Different Approaches for Detecting Emotion from Text," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1110, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1757-899X/1110/1/012009.

[2] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.

[3] K. Shrivastava, S. Kumar, and D. K. Jain, "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network," *Multimed. Tools Appl.*, vol. 78, no. 20, pp. 29607–29639, Oct. 2019, doi: 10.1007/s11042-019-07813-9.

[4] F. Anzum and M. L. Gavrilova, "Emotion Detection From Micro-Blogs Using Novel Input Representation," *IEEE Access*, vol. 11, pp. 19512–19522, 2023, doi: 10.1109/ACCESS.2023.3248506.

[5] Y. Qian, W. Liu, and J. Huang, "A Self-Attentive Convolutional Neural Networks for Emotion Classification on User-Generated Contents," *IEEE Access*, vol. 8, pp. 154198–154208, 2020, doi: 10.1109/ACCESS.2019.2938560.

[6] A. Rusli, A. Suryadibrata, S. B. Nusantara, and J. C. Young, "A Comparison of Traditional Machine Learning Approaches for Supervised Feedback Classification in Bahasa Indonesia," *IJNMT Int. J. New Media Technol.*, vol. 7, no. 1, pp. 28–32, Jul. 2020, doi: 10.31937/ijnmt.v1i1.1485.

[7] G. P. Wiratama and A. Rusli, "Sentiment Analysis of Application User Feedback in Bahasa Indonesia Using Multinomial Naive Bayes," in *2019 5th International Conference on New Media Studies (CONMEDIA)*, Bali, Indonesia: IEEE, Oct. 2019, pp. 223–227. doi: 10.1109/CONMEDIA46929.2019.8981850.

[8] H. Zhang, A. Jolfaei, and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019, doi: 10.1109/ACCESS.2019.2949741.

[9] M. N. Meqdad, F. Abdali-Mohammadi, and S. Kadry, "Recognizing emotional state of user based on learning method and conceptual memories," *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 18, no. 6, p. 3033, Dec. 2020, doi: 10.12928/telkomnika.v18i6.16756.

[10] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "Ravdness, Crema-D, Tess Based Algorithm for Emotion Recognition Using Speech," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India: IEEE, Jan. 2022, pp. 1625–1631. doi: 10.1109/ICSSIT53264.2022.9716313.

[11] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern Recognit. Lett.*, vol. 93, pp. 133–142, Jul. 2017, doi: 10.1016/j.patrec.2016.12.009.

[12] D. Seal, U. K. Roy, and R. Basak, "Sentence-Level Emotion Detection from Text Based on Semantic Rules," in *Information and Communication Technology for Sustainable Development*, M. Tuba, S. Akashe, and A. Joshi, Eds., in Advances in Intelligent Systems and Computing, vol. 933. Singapore: Springer Singapore, 2020, pp. 423–430. doi: 10.1007/978-981-13-7166-0_42.

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[14] L. Singh, S. Singh, and N. Aggarwal, "Two-Stage Text Feature Selection Method for Human Emotion Recognition," in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, C. R. Krishna, M. Dutta, and R. Kumar, Eds., in Lecture Notes in Networks and Systems, vol. 46. Singapore: Springer Singapore, 2019, pp. 531–538. doi: 10.1007/978-981-13-1217-5_51.

[15] J. Pang *et al.*, "Fast Supervised Topic Models for Short Text Emotion Detection," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 815–828, Feb. 2021, doi: 10.1109/TCYB.2019.2940520.

[16] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," *J. Intell. Syst.*, vol. 31, no. 1, pp. 113–126, Jan. 2022, doi: 10.1515/jisys-2022-0001.

[17] R. Liu, Y. Shi, C. Ji, and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," *IEEE Access*, vol. 7, pp. 85401–85412, 2019, doi: 10.1109/ACCESS.2019.2925059.

[18] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning.," *J. Pers. Soc. Psychol.*, vol. 66, no. 2, pp. 310–328, 1994, doi: 10.1037/0022-3514.66.2.310.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/ARXIV.1810.04805.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv, Feb. 29, 2020. Accessed: Jan. 08, 2023. [Online]. Available: http://arxiv.org/abs/1910.01108

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, doi: 10.48550/ARXIV.1412.6980.