Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.7 No.3

Designing SBERT and Ensemble Learning Mixture Model for Sentiment Analysis of Children with Developmental Disabilities

Hyeonji Kim[1)] and Yoosoo Oh[2,*)]
[1)]Dept.of Information and Communication Engineering
, Daegu University, Korea
[2)]School of AI, Daegu University, Korea

Abstract: Children with developmental disabilities communicate through speech and actions. However, children with developmental disabilities have difficulty communicating their feelings to others using language. Nevertheless, children with developmental disabilities express their emotions primarily through words rather than actions. Therefore, this paper proposes a mixing model that combines SEBRT, a language embedding model, and Voting Classifier, a machine learning ensemble model. The proposed mixture model can improve the performance of sentiment analysis models for developmentally disabled children by not only sentence embedding but also by giving them semantic weights.

Keywords: Developmentally disabled children, NLP, SBERT, Machine Learning, Ensemble Learning

1. Introduction

Developmental disabilities are characterized by delays and abnormalities in social relationships, communication, and cognitive development and refer to an age-unsuitable physical and mental developmental disorder. Developmental disorders are characterized by delayed or absent development of language and communication. In developmental disabilities, children have difficulty communicating their emotions to others using language, facial expressions, or gestures [1] . Among developmental disorders, autism shows persistent impairment in social communication and interaction. People with autism most often use language to express their emotions. Emotional sentences of

[*]Corresponding author: yoosoo.oh@daegu.ac.kr
Received: Aug 3, 2023; Accepted: Sep 11, 2023; Published: Sep 30, 2023

children with autism disorder are characterized by the combination or repeated use of specific words.

There are language and machine learning classification models for identifying language that includes emotions. Among language models, SBERT (Sentence-BERT) is a model in which sentence embedding of BERT is improved through fine-tuning. SBERT derives a fixed-size vector for input sentences and can use cosine similarity to derive semantically meaningful sentence embeddings [3].

Among machine learning models, the Voting Classifier ensemble model derives the optimal result by voting the learning results of various machine learning models [9]. As a result, ensemble models yield improved performance over single machine learning models [9].

This paper proposes the S-BEL (Sentence-Bert Ensemble Learning) Mixture model, which combines the SBERT and ensemble models. The proposed S-BEL Mixture model derives fixed vectors through sentence embedding and assigns semantic weights to sentences. The proposed S-BEL Mixture model can generate answers in the form of appropriate chatbots that correspond to emotion analysis and emotions of children with autism who express emotions with specific words.

2. Related Research

Park Sang-min, Lee Jae-Yoon, Sung Yuri, and Kim Jae-Eun proposed an SBERT model learning method that can express semantic vectors considering the amount of keyword information. Park Sang-min, Lee Jae-Yoon, Sung Yuri, and Kim Jae-Eun constructed {positive and negative} keyword training data by sentence n-gram to construct training data containing keyword information. Park Sang-min, Lee Jae-Yoon, Sung Yuri, and Kim Jae-Eun used the SBERT model to learn the data from the collected learning sentences and constructed keyword pairs. In this related study, an SBERT model trained on data reflecting keywords showed a 2.74% improvement in performance compared to the existing SBERT model [2]. Yoon Hye-jin, Ku Ja-hwan, and Kim Eung-mo used five types of word embedding techniques and three types of machine learning classification models to advance sentiment classification. CounterVectorizer, TfidfVectorizer, Word2vec (CBOW), Word2vec (Skip-gram), and Pretrain_Word2vec were used for five-word embeddings, and Decision Tree, RandomForest, and Logistic Regression models were used as machine learning classification models. As a result of testing, related research showed that CounterVectorizer outperforms TfidifVectorizer [4].

Park Sang-min, Lee Jae-Yoon, Son Yu-ri, and Kim Jae-Eun created two keywords through n-gram and conducted learning by pairing keywords and sentences. Then, Yoon Hye-jin, Koo Ja-hwan, and Kim Eung-mo applied three types of machine learning to analyze sentiment proceeded. Since emotion models are specific and diverse, better performance can be obtained by combining each machine learning rather than using a single model to improve the accuracy of emotion machine learning. Therefore, in this paper, keywords are subdivided into six emotions, and learning is conducted. In addition, this paper proposes a model that combines several machine learning models using a natural language processing model and an ensemble technique.

3. S-BEL Mixture Model

This paper constructed a model using more than 10,000 emotional datasets provided by the National Information Society Agency. The National Information Society Agency's emotional data has six major categories: emotions, human response, and system response. In this paper, we use human response 1, system response 1, and significant categories of emotions ("joy," "embarrassment," "anger," "anxiety," "wound," and "sadness") as features.

The overall diagram of the model proposed in this paper is shown in Figure 1 below. The proposed model consists of a SentenceBERT model, a machine learning classification model, and an S-BEL Mixture model.

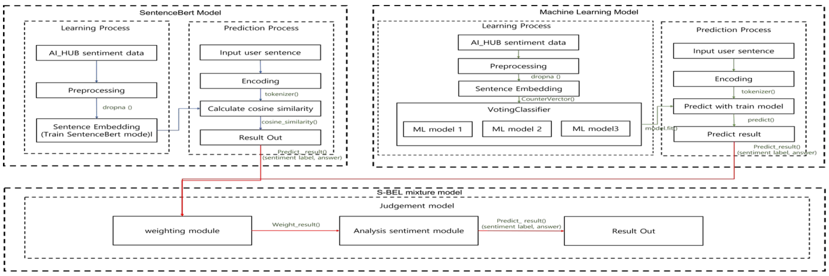

Figure 1. S-BEL Mixture Model Diagram
(ML1: MultinomialNB, ML2 : RandomForest, ML3 : XGBoost)

The S-BEL Mixture model proposed in this paper gives weights using the prediction results of the SBERT model and the ensemble learning model. The SentenceBERT model is a pre-trained model with a large-capacity Korean language model. In this paper, we learn the finely tuned SentenceBERT model using the data consisting of human sentence one and emotion_large category pairs as feature values. The learned SentenceBERT proceeds with an encoding that vectorizes the sentences entered by the user to make predictions corresponding to user input. Sentences vectorized by encoding predict the closest sentence and emotion by cosine similarity calculation with the trained model. Cosine similarity is calculated as Equation 1 [8].

$$\textbf{similarity} = \cos(\theta) = \frac{A * B}{\| A \| \| B \|}$$

Equation 1. Cosine Similarity Formula

Ensemble learning, one of machine learning classification, combines the prediction results of different machine learning prediction models to derive a single final prediction result. The proposed model uses XGBoost, RandomForest, and MultinomialNB classification models to construct the VotingClassifer model. A Hard-Voting model finalizes the predicted value of the class that gets the most votes among the results of each classifier model by majority classification. In contrast, a Soft-Voting model predicts the class with the highest average probability for each classifier prediction exists.[7] The proposed model was modeled using the Soft-Voting model. In this paper, sentences were vectorized by the number of occurrences of words through CounterVectorizer for collected emotional data. Afterward, we trained the MultinomialNB, RandomForest, and XGBoost models to the VotingClassifier model through the pipeline. The proposed ensemble model proceeds with embedding user input sentences to predict emotion corresponding to user input. The emotional result of the input sentence is output through the model's prediction according to the probability.

The weighting formula for the S-BEL Mixture model proposed in this paper is shown in Equation 2 below. In the proposed model, b_s is the similarity measure of the SBERT model, W_ms is the accuracy value of the ensemble learning model, and X_mp is the probability value measured by the ensemble learning model. Equation 2 performed a correlation analysis of accuracy and prediction probability via the product of W_ms and X_mp. In addition, equation 2 shows the deflection value via the value of b_s−1. Furthermore, Equation 3 applied a sigmoid function formula for scaling between 0 and 1 [10].

$$S.BEL_w = W_{ms} * X_{mp} + (b_s - 1)$$

Equation 2. S-BEL Mixture Model Weight Formula ($S.BEL_w$: Weight result value, W_{ms}: $EL accuracy value of the model$, X_{mp} : Probability value measured through EL model, b_s: $Similarity\ measure\ of\ SBERT\ model$)

$$S.BEL_w = \frac{1}{e^{-(S.BEL_w)}}$$

Equation 3. Weight scaling formula of S-BEL Mixture model

The proposed S-BEL Mixture model checks whether the S-BEL Mixture weight value is above or below a threshold through a determination module. If the judgment result is above the threshold, an emotion label and chatbot-type answer corresponding to the emotion predicted by the proposed model are output. Conversely, if the determination result is less than the threshold value, it is determined that the prediction result is wrong, and the result value is not output.

4. Experiments and Results

This paper compares the accuracy of the classification model and the proposed model, as shown in Fig.2. In this paper, accuracy measurements were performed using a method of dividing the train data and the test dataset in a 7:3 ratio and dividing the total number of samples by the correctly predicted number of samples. As a result of the experiment, the accuracy was higher when progressing using ensemble learning than when progressing with sentiment analysis using each machine learning algorithm. Furthermore, it was confirmed that the proposed S-BEL Mixture model achieved the highest accuracy among the compared models, with an accuracy of about 92% when the sentiment analysis was performed.

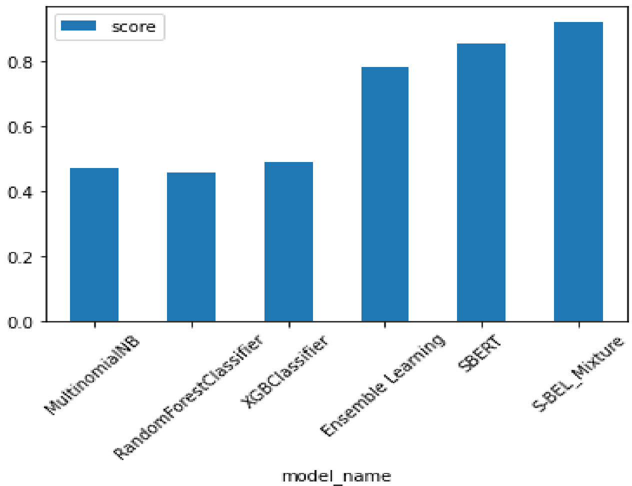

Figure 2. Benchmark Model Accuracy Comparison

In this paper, experiments were conducted on five sentences, as shown in Table 1. In this paper, the experiment proceeded through random input sentences. In this paper, we compare the actual emotion labels for random input sentences with the emotion labels predicted by the proposed mixture model. In addition, an experiment to see if the chatbot-type answers presented by the proposed model were appropriate.

Table 1. EXPERIENCE RESULT (EXCERPTS OF SOME SENTENCES FROM THE TEST DATA)

	Input Sentence	Real sentiment labels	Predict sentiment labels	Answers in the form of chatbots
1	Sad	Sadness	Sadness	You must be sad.
2	That's very annoying.	Anger	Anger	You're annoying.
3	I'm angry.	Anger	Anger	I think you'll be so angry.
4	Glad.	Joy	Joy	Oh, you must be so happy!
5	I don't know what to do after I retire.	Embarrassment	x	x

As a result of the experiment, as shown in Table 1, it can be confirmed that the actual emotion label and the predicted emotion label for the input sentence match at a rate of about 85%. In addition, the proposed model matched the actual emotion label of simple sentences such as 'I am angry' and 'I am happy' with the predicted emotion label. The proposed model can be confirmed to output appropriate chatbot-style answers to the predicted sentiment analysis.

5. Result

In this paper, we propose an emotion analysis model using SBERT, ensemble machine learning, and S-BEL Mixture weights. The results of emotion analysis through the proposed S-BEL Mixture model were found to have the highest accuracy compared to other emotion analysis models.

The currently proposed sentiment analysis model used the sentiment data of ordinary people rather than the dialogue data of autistic children for sentiment analysis. Future research plans to learn by collecting dialogue data on children with autism. SBERT can be fine-tuned according to the training data. Therefore, future research will improve the performance of learning models by optimal fine-tuning. In future research, in addition to models such as RandomForest and XGBoost used to improve the accuracy of the VotingClassifier model, we plan to combine various machine learning models to present the optimal combination and proceed with learning. We also plan to extend the proposed mixture model to improve its performance of the model.

References

[1] JeongMi Lee. "Study on Spontaneous Communicative Behavior in Children with Autism." Emotional and Behavioral Disorder Research 23, no.4 (2007): 345-359.

[2] SangMin Park, JaeYoon LeeYu Ri Sung, JaeEun Kim. (2022). SBERT-IQ: A Sentence-BERT-based embedding model considering keyword information content. Proceedings of the Korean Information Science Society, (), 1058-1060.

[3] Nils Reimers, Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," In Proceedings of Empirical Methods in Natural Language Processing, pp. 3980-3990, 2019.

[4] HyeJin Yoon, JahWan Koo, and UngMo Kim. "Comparison of Word Embedding Performance for Sentiment Classification." Proceedings of the Annual Meeting of the Korean Information Processing Society 28.2 (2021): 760-763.

[5] Y. J. Jang, H. J. Kim, and H. J. Jo, "Data Mining," KNOU PRESS, 2016, 1-200

[6] W. Yun, D.H. Seo, S. Min, H. Nam, "Breast Cancer Classification using RandomForest and XGBoost," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 113-114, 2021.

[7] M. C. Jeong, J. Lee, and H. Oh, "Ensemble Machine Learning Model Based YouTube Spam Comment Detection," Journal of the Korea Institute of Information and Communication Engineering, vol. 24, no. 5, pp. 576–583, May 2020.

[8] SungEun Gil. (2020). Research on action-based authentication by cosine similarity measure Proceedings of the Annual Meeting of the Korean Information Processing Society, 27(1), 165-168.

[9] HuiWon Song, JuHyun Lee, SungHae Jun, SangSung Park. (2022). A Study on Inferable Patent Evaluation Model using Ensemble Method. Journal of Korean Institute of Intelligent Systems, 32(5), 379-385.

[10] Sanjeev Kumar, Mahesh Chandra). "Detection of Microcalcification Using the Wavelet-Based Adaptive Sigmoid Function and Neural Network." Journal of information processing systems 13.4 (2017): 703-715.