Toward Ubiquitous Foreign Language Learning Anxiety Detection

Daneih Ismail[1][0000−0003−4975−3014] and Peter Hastings[1][0000−0002−0183−001X]

DePaul University, Chicago IL 60614, USA
dismail1@depaul.edu
peterh@cdm.depaul.edu

Abstract. We present a novel design for detecting Foreign Language Anxiety (FLA) while the learner is using an English as a second language system (ESL). Our method uses sensor-free metrics to avoid disrupting the learning process. We evaluated the validity and reliability of several machine learning models using data from two different systems. Using 9 features extracted from the interaction, we found that Random Forest, XGBoost, and Gradient Boosting Regressor provided suitably accurate predictions of anxiety, and outperformed Linear Regression, Support Vector Regressor, and Bayesian Ridge Regression.

Keywords: Affect detection · FLA· ML · Sensor-free metrics

1 Introduction

Automated detectors for predicting emotions such as engagement, boredom, confusion, and frustration have achieved high accuracy [1]. However, there is still a need to improve prediction of Foreign Language Anxiety (FLA), a significant impediment to learners of new languages [7]. Previous emotion research has shown that multiple factors affect learners' vulnerability to FLA, such as task complexity [3], academic achievement, gender and age [10]. Horwitz, et al [4] found that there are three dominant components that influence FLA: fear of negative evaluation, communication apprehension, and test anxiety [4]. Based on this, they developed a well-established and validated instrument for measuring FLA, the Foreign Language Classroom Anxiety Scale (FLCAS) [4]. The FLCAS was developed for use in a classroom context, but it has also been shown to correlate well with self-reported anxiety within online tutoring system [6].

Sensor-free metrics detect emotions from the users' interactions with the system without using any physical monitors [7]. Previous researchers have built sensor-free emotion detector, comparing various machine learning algorithms to reach the best prediction [1]. In this work, we focused on Foreign Language Anxiety in particular. We extracted features from learners' interactions with ESL systems then built machine learning models to predict FLA from those features. We used self-reports of FLA as ground truth for our predictions. We analyzed the following research questions:

Research Question 1: "Can FLA be detected without learning interruption when using ESL learning systems?" This question has two sub-questions:

RQ1A: What features best predict FLA?

RQ1B: What machine learning methods are better for predicting FLA?

Research Question 2: "How well can sensor-free detectors be generalized to other emotionally intelligent foreign-language/ESL learning systems?"

2 Method

The data was collected from two different ESL systems. Dataset 1 came from a system focused on practice, with no tutorial and no scoring. Thirty participants did 27 exercises, covering vocabulary, grammar, listening, conversation, and speaking, providing data from a total of 810 exercises. Dataset 2 came from an online system which included video tutorials and feedback on the answers [8]. 29 participants did 26 exercises which covered vocabulary, grammar, listening, reading, and writing, producing data from a total of 704 exercises. For both experiments the participants completed level of anxiety self-report after each exercise.

From each of the exercises, 16 features were extracted. Following [7], we used the average of the three FLCAS component scores: fear of negative evaluations, communication apprehension, and test anxiety [4]. Following [10], we included the participant's age, gender, education level, English level, exercise score, duration, exercise topic, score on the preceding exercise, the percentage of previous incorrect scores, the percentage of previous correct scores, average percentage of all previous exercises, and average duration of exercises of the same section. We did a correlation analysis and set an absolute threshold value of 0.5 to eliminate multicollinearity. Then we used the Gini importance feature selection algorithm to distill the features that could cause overfitting and kept only the features that improved the model. From an original set of 16 features, we ended up with 9 features that provided an acceptable accuracy with the least bias.

We made predictions using regression instead of classification because we used continuous-valued self-report to measure anxiety in order to capture moment-to-moment emotion fluctuation [9] and to provide more accurate high-resolution measurements (as opposed to, e.g., the Likert classification scale). The methods we evaluated were Random Forests, XGBoost, Gradient Boosting Regressor, Linear Regression, Bayesian Ridge Regression, and Support Vector Regressor. We implemented these machine learning models in the scikit-learn library in Python. We evaluated each detector using 10-fold cross-validation.

3 Results

Regarding RQ1A, determining which features are predictive of FLA, based on the correlation analysis and Gini importance algorithm, the final set of features that reliably predicted FLA were: exercise score, percentage of all previous exercise scores, percentage of previous incorrect scores, exercise duration, relevant

exercise duration, FLCAS score, English level, exercise topic, and the participant's age. Gini indicated that the most important features were FLCAS score followed by the average percentage of all previous exercises.

With respect to RQ1B, on the types of machine learning methods, for Dataset 1, the Random Forest method was most accurate, predicting 47% of the variance of FLA. XGBoost was close behind, predicting 45%. For Dataset 2, both methods predicted 66% of FLA. The performance of Gradient Boosting was slightly behind that of the other ensemble methods. In contrast, the non-ensemble methods performed much worse, predicting a maximum of 21% of the variance of FLA. For RQ2, which focused on the robustness of these features and models across the different systems and datasets, we found that the set of most important features for both datasets was identical, with slight differences in ranking. The relative performance of the models was identical across the datasets.

4 Discussion and Conclusions

Prior research used FLCAS components and exercise score as sensor-free metrics to predict FLA [7]. Here, we extend this by uncovering features that produce better predictions using machine learning without interrupting the learning process. Previous research found that FLA could be predicted up to 43% using Linear Regression using FLCAS components and exercise scores, but only by including self-reports of system and language difficulty after each exercise. Without these intrusive self-report measures, the maximum prediction was 20%. Here, we found that by augmenting the FLCAS scores with behavioral features from the participants' interactions with the systems, and by using machine learning models, we can predict up to 66% of the learner's anxiety without interruptions. This level of accuracy is imperfect yet satisfactory; affect detection is extremely difficult because it is not directly accessible [1].

Because the two datasets had identical highest feature importance rankings, these features should provide reliable predictive performance for any e-learning system that teaches English as a foreign language. They are also easily derived from pretest and behavioral data.

As mentioned above, previous research found that sensor-free metrics could predict FLA using Linear Regression, accounting for 20% of the variation in anxiety [7]. Here, we found that machine learning models can produce much more accurate predictions. It is clear that ensemble learning models (Random Forest, XGBoost, Gradient Boosting Regressor) outperform non-ensemble models (Linear Regression, Bayesian Ridge, and Support Vector Regressor). The high performance of the ensemble learning models is consistent with other research demonstrating the robustness, reliability, and stability of these methods [5]. Because these ensemble learning methods can produce acceptably accurate predictions of the learner's level of anxiety, they can be used to support an emotionally intelligent tutoring system which can adaptively provide interventions according to the learner's current emotional state.

When the performance of a model on a second dataset is the same or better than on the one for which it was developed, that provides evidence for the reliability of the model [2]. Here, we extracted a set of features from one system and dataset, evaluating predictive performance with multiple models. Then, using the same features, found that the relative performance was equivalent across the systems, and that all of the models actually produced better performance in the second dataset. Thus, we demonstrated the generalizability of this approach to any ESL system because the models used features that can be easily extracted from any such system. A limitation of this approach is that the features were selected along with the machine learning algorithm. For our future work we will build an emotionally intelligent tutoring system to detect FLA and reduce it by adaptively providing appropriate feedback, creating a more positive and effective learning environment.

References

1. Baker, R., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L.: Towards sensor-free affect detection in Cognitive Tutor Algebra. Proceedings of the Fifth International Conference on Educational Data Mining pp. 126–133 (2012)
2. Bosnić, Z., Kononenko, I.: An overview of advances in reliability estimation of individual predictions in machine learning. Intelligent Data Analysis **13**(2), 385–401 (2009)
3. Hashemi, M.: Language stress and anxiety among the English language learners. Procedia - Social and Behavioral Sciences **30**, 1811–1816 (2011)
4. Horwitz, E.K., Horwitz, M.B., Cope, J.: Foreign language classroom anxiety. The Modern Language Journal **70**(2), 125–132 (1986)
5. Hueniken, K., Somé, N.H., Abdelhack, M., Taylor, G., Marshall, T.E., Wickens, C.M., Hamilton, H.A., Wells, S., Felsky, D., et al.: Machine learning–based predictive modeling of anxiety and depressive symptoms during 8 months of the COVID-19 global pandemic: Repeated cross-sectional survey study. JMIR mental health **8**(11), e32876 (2021)
6. Ismail, D., Hastings, P.: Identifying anxiety when learning a second language using e-learning system. In: Proceedings of the 2019 Conference on Interfaces and Human Computer Interaction. pp. 131–140 (2019)
7. Ismail, D., Hastings, P.: A sensor-lite anxiety detector for foreign language learning. In: Proceedings of the 2020 Conference on Interfaces and Human Computer Interaction. pp. 19–26 (2020)
8. Ismail, D., Hastings, P.: Way to go! effects of motivational support and agents on reducing foreign language anxiety. In: International Conference on Artificial Intelligence in Education. pp. 202–207. Springer (2021)
9. Lottridge, D., Chignell, M., Jovicic, A.: Affective interaction: understanding, evaluating, and designing for human emotion. Reviews of Human Factors and Ergonomics **7**(1), 197–217 (2011)
10. Onwuegbuzie, A.J., Bailey, P., Daley, C.E.: Factors associated with foreign language anxiety. Applied Psycholinguistics **20**(2), 217–239 (1999)