Predicting Probability of Investment Based on Investor's Facial Expression in a Startup Funding Pitch

Arya Tri Prabawa[1], Merel M. Jung[1][0000−0002−3645−9425], Kostas Stoitsas[1], Werner Liebregts[2], and Itır Önal Ertuğrul[3]

[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands aryaprabawa1987@gmail.com, m.m.jung@tilburguniversity.edu, kstoitsas@gmail.com

[2] Jheronimus Academy of Data Science, St. Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands w.j.liebregts@tilburguniversity.edu

[3] Department of Information and Computing Sciences, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands i.onalertugrul@uu.nl

Abstract. Presenting an idea is a critical social interaction, especially in a startup funding pitch setting where initial investment is at stake. Understanding a listener's facial expression can then become extremely valuable in informing the level of engagement reached by the presenter. Predicting engagement level in other settings, such as an online study environment, has been explored in previous research, but none have explored to what extent an investor's facial expression can predict the investor's engagement during a funding pitch and in return predict the investor's probability to invest. In this study, we propose to use Long Short-Term Memory (LSTM) networks along with facial action units (AUs), facial features extracted with Convolutional Neural Networks (CNN), and the combination of both as features for automated prediction of probability of investment. The results show a promising prospect for the proposed LSTM models. Models using CNN features or combined AU and CNN features outperformed the AU-only model.

Keywords: Facial action units · Deep facial feature extraction · Entrepreneurial pitches · Long short-term memory networks.

1 Introduction

Communicating an idea is a delicate, yet ubiquitous social interaction. Getting an audience to understand and accept an idea requires not only an understandable and acceptable idea, but also an understandable and acceptable presentation. A startup funding pitch is one example where presenting an idea becomes a critical point by which a startup can secure its initial investment [4].

A previous study used computer vision and machine learning techniques to predict an investor's decision based on a startup founder's facial expression [8], but no previous studies have focused on the faces of the listeners: the investor's

facial expression as the audience. Meanwhile, in other settings such as online learning and advertising, studies have been conducted to predict acceptance based on facial expressions [9, 15]. Based on this gap, this study will investigate the predictive performance of an investor's facial expression during a startup funding pitch on their decision to invest.

Prediction and further investigation of investment decisions based on an investor's facial expression is important to understand the visual feedback of social interaction on the other end of a presentation. A better understanding of how well an audience responds to an idea being communicated can lead to more transparency in emotion, providing a better guess of what they would do or decide next [2]. Specific to the startup funding pitch setting, startup founders with big ideas can further improve their pitch delivery based on facial expression feedback from their investing counterpart, making sure that what they are trying to convey is fully understood. Understanding the way information is being perceived by an audience can also help improve communication in various media. Communication evaluation can then be implemented both in real time and without the biased nature of self-reported surveys [11].

Facial expressions can be captured from video recordings in various forms: facial Action Units (AUs) [7], facial landmark positions [20], and by using deep learning models such as Convolutional Neural Networks (CNNs) to extract features directly from the videos [10]. In terms of prediction methods based on sequential data, Recurrent Neural Networks (RNNs), such as the Long Short-Term Memory (LSTM) networks, have been reported as having better predictive performance compared to conventional machine learning methods [12].

In this study we evaluated the performance of LSTM regression models with three different sets of input features to predict the probability that investors will invest in a startup idea. Facial features were extracted from video recordings of the investors while listening to pitches considering either 1) AUs; 2) CNN features; 3) the combination of both features. Our models showed promising performances for predicting investment decisions based on investor's facial expressions with the models using CNN features or combining AU with CNN features performing best.

2 Related Work

2.1 Facial expressions as predictor of engagement

Investor's engagement has been proposed as a possible mediator for the effect of displayed entrepreneurial passion on investment in a startup funding pitch [16]. The engagement of the investors was measured through neural activity using functional Magnetic Resonance Imaging (fMRI) while listening to a startup founder's pitch. The result showed not only that higher levels of displayed entrepreneurial passion increased investor's engagement, but also that higher levels of engagement increased the investor's interest to invest. Engagement detection methods can be divided into 3 main categories: manual, semi-automatic and automatic [6]. Manual engagement detection methods require direct actions from

the subjects, such as using self-report surveys which require honesty, accurate self-perception, and time investment. Although semi-automatic methods helped to remove most of these biases through indirectly inferring a subject's actions, such as measuring response time on a certain task, it was outgrown by automatic methods in popularity. Besides using neural sensors to automatically detect engagement, computer vision methods also show increasing promise.

One of the ways computer vision can be used to infer engagement is detecting movement of facial landmarks. For example, the position of facial landmarks such as the eyes, nose, and mouth have been used to detect subjects' engagement while playing a game and listening to video instructions [20]. The facial landmarks position data was used to train three different machine learning methods: GentleBoost, Support Vector Machines (SVM) and Multinomial Logistic Regression (MLR). Results showed that machine learning methods, especially SVM and MLR, yielded comparable accuracy to human annotators in detecting engagement. In a study on engagement in video advertisements, eye movements and facial landmarks positions were used to train a dynamic model detecting the emotions of joy and surprise and the results showed a significant influence of expressed joy and surprise on the audience's engagement [18].

Another way to detect facial expressions is by using the AUs of the Facial Action Coding System (FACS) [7]. FACS is an anatomically-based system which looks at how certain movements in facial muscles can represent different facial expressions. Forty-four facial AUs were defined which can activate singularly or in combination with one another to display certain facial expressions. Isolating these AUs is very important for capturing facial expressions to analyze human emotions [19]. Facial AUs have been used for binary classification of student's engagement in a classroom setting [15]. Three variations of machine learning models (SVM, naive Bayes and random forest) were trained using the facial expressions of 123 students. The presence of the following AUs were used as input to train the model: additive combination of AU7 (Lid Tighter) and AU12 (Lip Corner Puller), AU5 (Upper Lid Raiser), AU25 (Lips Part) and AU26 (Jaw Drop). The results showed that the SVM model performed best, with an F1-score of 0.861 on 10-fold cross-validation. Another study showed that higher levels of interest correlated with higher intensities of AU6 (Cheek Raiser) combined with AU12 (Lip Corner Puller) which together form an enjoyment smile and also with AU7 (Lid Tightener) which is linked to the expression of attention [17].

An alternative to hand-crafted facial expression features such as AUs, is the use of deep neural networks such as CNN models to extract facial features directly from images or videos. For experiments with small datasets, pre-trained CNN models are usually employed. For example, [10] proposed the VGG-Face pre-trained CNN model to estimate facial expression intensity. In this regression task, VGG-Face outperformed the use of hand-crafted facial landmarks features.

As previous studies have shown that facial expressions can be used for engagement detection in the domains such as online learning and advertisement, we propose to apply such an approach to an entrepreneurial context. To our knowledge, no previous research has focused on an investor's facial expressions

to predict their probability to invest. Facial expressions will be captured from video recordings of investors in the form of AUs and facial expressions extracted by a CNN to compare the individual and combined performance of both feature extraction methods.

2.2 Long Short-Term Memory (LSTM) networks for predicting engagement

Looking at engagement as a signal which occurs over time, previous studies have considered sequential models to better capture facial expressions for predicting engagement. For example, Long Short-Term Memory (LSTM) networks were used to predict engagement based on facial expressions from novice-experts interactions [5]. Engagement was predicted based on AUs extracted by OpenFace [1] as well as head rotation and gaze detection. The results showed that LSTM outperformed other non-sequential machine learning models, such as naive Bayes, decision trees and conventional neural networks. Facial AUs were shown to have a higher contribution in engagement prediction compared to head rotation and gaze detection. In a slightly different study, a LSTM model was used to extract facial expressions for emotion prediction showing that the sequential nature of the models outperformed non-sequential methods [12]. In a multimodal approach facial AUs, head pose, gaze detection, facial landmarks positions, and body posture movements were used as inputs for the LSTM model [21]. Down-sampling was implemented to reduce computation time since minimal movement was expected. The results showed that a more conservative LSTM model with fewer units yielded better prediction results, overcoming the risks of overfitting on a small training set.

These previous studies showed the promise of using LSTM models to better capture the sequential nature of facial expressions for the prediction of engagement. Therefore, LSTM models will be implemented in this study for detecting engagement based on investors facial expressions to predict their probability to invest in a startup funding pitch.

3 Method

3.1 Entrepreneurial pitch dataset

The dataset used in this study contains video recordings from entrepreneurial pitch competitions and survey data [13]. The data collection has been approved by the ethics committee of Tilburg University and written informed consent was obtained for participation. Pitchers were university students who participated in the pitch competition as part of a course on entrepreneurship in data science. The pitches were evaluated by a panel of three investors who were all professionals with extensive experience in the industry. Pitchers had a maximum of three minutes to pitch their start-up idea on behalf of their group followed by a question and answering round. After each pitch, the investors were asked to complete

a questionnaire evaluating the pitches including a rating of the probability that they would invest in the business idea. In total, across four pitch competitions, there are 87 recordings from investors evaluating 29 pitches. The investor panel members differed for each of the four pitch competitions whereas within each competition the panel was the same. As a result each investor evaluated multiple pitches. For the purpose of this study, we focused on the first minute of the video recordings (1080p resolution at 25 FPS) of investors listening to the pitches to investigate whether investment decisions can be predicted based on first impressions.

3.2 Comparison of investment prediction models

In this study we compared three regression models for predicting the probability that investors would invest in a startup idea on a scale from 0-100 using the questionnaire ratings as ground truth. Three models were trained to evaluate the performance of different facial features:

- AU+LSTM in which sequences of AU intensity values were fed to the LSTM.
- CNN+LSTM in which sequences of CNN features were fed to the LSTM.
- AU+CNN+LSTM in which sequences of AU intensity values and CNN features were concatenated and then were fed to the LSTM.

3.3 Preprocessing

The following preprocessing steps have been implemented: a frame selection method was be applied to shorten the time sequences fed into the models, facial AU data was normalized, and oversampling was implemented for training all models.

Frame selection RNNs such as LSTM are suboptimal when used for modeling long sequences [14]. Therefore, the video recordings of the first minute of the pitch were down-sampled to 1 frame per second resulting in a sequence of 60 frames.

Normalization In a prediction task with AUs as input, a squared min-max normalization has been found to be useful to optimize the training process [22]. Min-max normalization prevented the values from being too different or too similar. Squaring the min-max normalized values then emphasized both the rewards and punishments for relevant and irrelevant features respectively. Formula 1 below shows the calculation of the squared min-max normalization.

$$AU'_x = \left(\frac{AU_x - min\{AU_i\}}{max\{AU_i\} - min\{AU_i\}}\right)^2 \quad (1)$$

Here, AU_x represents each AU while AU_i represents all AUs.

Random over-sampling As can be seen in Figure 1, there is an imbalanced distribution of the probability to invest ratings in the dataset, with more observations available for the middle-to-lower ratings. To account for this imbalance, random over-sampling has been applied to the training set. With random over-sampling, new samples for the underrepresented values are generated by randomly sampling with replacement from available data.

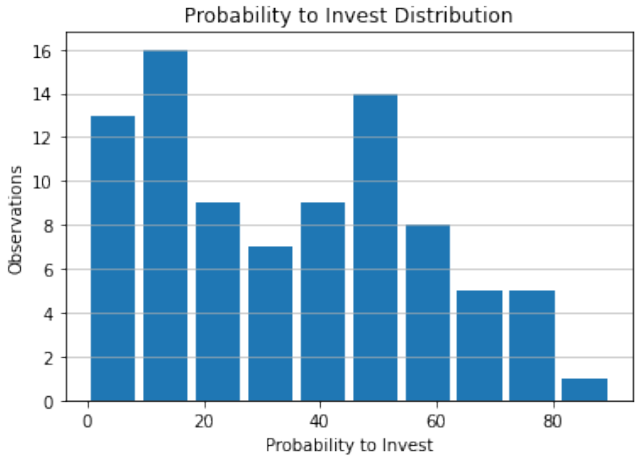

Fig. 1. Distribution of the probability to invest values across the dataset.

3.4 AU extraction

The intensities of each of the AUs listed in Table 1 were extracted per frame using OpenFace [1]. OpenFace extracts the intensities of AUs on a continuous scale from 0 to 5.

3.5 CNN facial feature extraction

Aligned faces from the investors extracted by OpenFace [1] were used as input (size: 112x112x3) for the CNN model. The model was an Xception network [3] with pre-trained ImageNet weights and global average pooling, and without a fully-connected layer at the top[4]. Image sequences were fed into the network using Keras's *TimeDistributed layer*[5]. Based on initial experimentation, only the last layer was retrained and 1024 features were extracted.

[4] `https://keras.io/api/applications/xception`
[5] `https://keras.io/api/layers/recurrent_layers/time_distributed`

Table 1. Facial Action Units (AUs) included in this study.

Features	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU7	Lid Tightener
AU9	Nose Wrinkler
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU23	Lip Tightener
AU25	Lips Part
AU26	Jaw Drop

3.6 LSTM networks

The facial features were used as input for the three LSTM models: AU+LSTM (16 features), CNN+LSTM (1024 features), AU+CNN+LSTM (1040 features). The models' architectures consisted of one LSTM layer on top of 1-4 dense layers. Sequences were fed into the LSTM using Keras's *TimeDistributed layer*. Models were optimized using the following parameters: recurrent units (8, 16, 32), dropout (0, 0.05, 0.1, 0.5), kernel regularizer (None, L1, L2, L1L2), dense layers (1, 2, 3, 4), dense units (8, 16, 32, 64, 128, 256, 512), optimizer (RMSprop, Adam, Nadam, SGD) and learning rate (0.001, 0.0001, 0.00001). The AU+LSTM model was trained for 800 epochs and the CNN+LSTM and AU+CNN+LSTM models were trained for 300 epochs. Early stopping was used for all models.

3.7 Training and evaluation

Data was divided into four folds, one for each of the four pitch competitions. Models are trained and evaluated using competition-independent 4-fold cross-validation using two folds for training, one for validation, and one for testing. This ensures model generalization as the data from a particular investor has never be used for both training and testing. The final regression performance metrics Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were calculated by averaging across all folds.

4 Results

Table 2 shows the MAE and RMSE scores of all three proposed investment prediction models. Based on the MAE scores, the best performing model is

the AU+CNN+LSTM model, followed by the CNN+LSTM model. In terms of RMSE score, however, the CNN+LSTM model slightly outperforms the AU+CNN +LSTM model. The AU+LSTM model showed the worst performance.

Table 2. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) scores for the three proposed models.

Model	MAE	RMSE
AU+LSTM	20.78	24.87
CNN+LSTM	19.62	**23.27**
AU+CNN+LSTM	**17.80**	23.61

A breakdown of the MAE score distribution across different levels of probability to invest ratings is shown in Figure 2. All three models show a similar pattern of more accurate predictions for the lower ratings. Although the AU+CNN+LSTM model is shown to outperform the CNN+LSTM and AU+LSTM models in some levels (e.g., 0-29% probability to invest) the CNN+LSTM and AU+LSTM models outperformed the AU+CNN+LSTM model on other levels (e.g., 30-59% probability to invest).

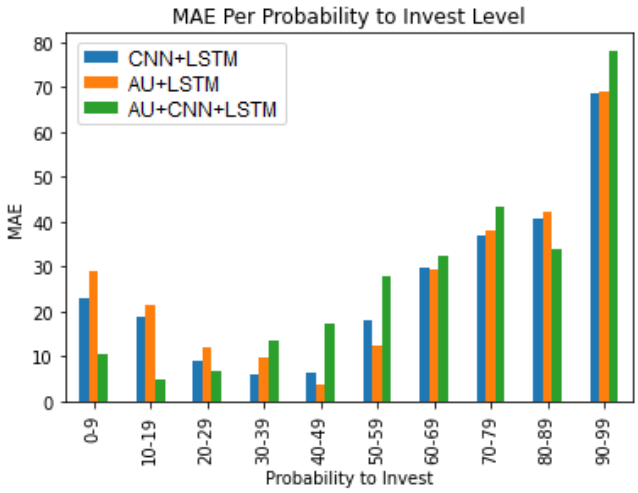

Fig. 2. Distribution of Mean Absolute Error (MAE) scores across models and probability to invest ratings.

5 Discussion

With the goal of providing better feedback on social interactions within a startup funding pitch setting, this study aims to answer the question of how well investor's facial expressions can predict the probability that they will invest. To answer this question, video recordings of the investors listening to pitches of startup ideas were used to develop prediction models. Three models were compared by feeding different feature sets (AUs, CNN extracted features, or a combination of both) into a LSTM network. The result showed promising results for the prediction of the probability to invest based on investor's facial expressions while judging startup pitches. Models using CNN features or a combination of AU and CNN features outperformed the model using AUs-only.

Previous studies have already shown promising results for the detection of engagement and interest based on facial expressions in various other settings [15, 17, 18, 20]. Our findings add to this body of research by providing valuable information on the use of listener's facial expressions as a predictor of engagement in the newly researched context of startup funding pitches. Especially since previous work found that higher levels of investor's engagement increased interest to invest [16].

A limitation of our current models is that despite leveraging oversampling techniques to counteract the imbalanced distribution of ratings, our models did not perform well for higher probability to invest ratings. Further studies can explore improvements to the prediction models by training on a larger dataset with a more balanced distribution of ratings. Moreover, a follow-up study will be required to investigate which facial features are most predictive of investment outcome. In addition, the predictive performance of facial expressions combined with other behavioral cues such as body expressions and eye gaze can be explored in a multimodal analysis to further understand the extent to which non-verbal behavior of investors is predictive of their decision to invest.

6 Conclusion

To conclude, the results of the current study show promising results for the detection of engagement based on investor's facial expressions for the prediction of the probability that they will invest in a startup idea. These findings can serve as a building block to apply predictive models in real-time to provide feedback to pitchers during pitching sessions. A continued effort in understanding the facial expressions of listeners can boost the quality of social interaction through various media and provide better feedback to those pitching their ideas.

References

1. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). pp. 59–66 (2018). https://doi.org/10.1109/FG.2018.00019

2. Barrett, L., Adolphs, R., Marsella, S., Martinez, A., Pollak, S.: Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest **20**, 1–68 (07 2019). https://doi.org/10.1177/1529100619832930
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
4. Clark, C.: The impact of entrepreneurs' oral 'pitch' presentation skills on business angels' initial screening investment decisions. Venture Capital **10**(3), 257 – 279 (2008). https://doi.org/10.1080/13691060802151945
5. Dermouche, S., Pelachaud, C.: Engagement modeling in dyadic interaction. pp. 440–445 (10 2019). https://doi.org/10.1145/3340555.3353765
6. Dewan, M., Murshed, M., Lin, F.: Engagement detection in online learning: a review. Smart Learning Environments **6** (01 2019). https://doi.org/10.1186/s40561-018-0080-z
7. Ekman, P., Friesen, W.V.: Facial action coding system: a technique for the measurement of facial movement (1978)
8. Hu, A., Ma, S.: Human interactions and financial investment: A video-based approach (2020)
9. Isabella, G., Vieira, V.A.: The effect of facial expression on emotional contagion and product evaluation in print advertising. RAUSP Management Journal **55**, 375–391 (2020)
10. Kawashima, T., Nomiya, H., Hochin, T.: Facial Expression Intensity Estimation Using Deep Convolutional Neural Network, p. 7–12. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3468081.3471060
11. Lewinski, P., Fransen, M., Tan, E.: Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. Journal of Neuroscience, Psychology, and Economics **7**, 1 (03 2014). https://doi.org/10.1037/npe0000012
12. Li, T.H.S., Kuo, P.H., Tsai, T.N., Luan, P.C.: Cnn and lstm based facial expression analysis model for a humanoid robot. IEEE Access **7**, 93998–94011 (2019). https://doi.org/10.1109/ACCESS.2019.2928364
13. Liebregts, W., Urbig, D., Jung, M.M.: Survey and video data regarding entrepreneurial pitches and investment decisions. Unpublished raw data (2018-2021)
14. Luo, Y., Chen, Z., Yoshioka, T.: Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. pp. 46–50 (05 2020). https://doi.org/10.1109/ICASSP40776.2020.9054266
15. Manseras, R., Eugenio, F., Palaoag, T.: Millennial filipino student engagement analyzer using facial feature classification. IOP Conference Series: Materials Science and Engineering **325**, 012006 (mar 2018). https://doi.org/10.1088/1757-899x/325/1/012006
16. Shane, S., Drover, W., Clingingsmith, D., Cerf, M.: Founder passion, neural engagement and informal investor interest in startup pitches: An fmri study. Journal of Business Venturing **35**(4), 105949 (2020). https://doi.org/10.1016/j.jbusvent.2019.105949
17. Soleymani, M., Mortillaro, M.: Behavioral and physiological responses to visual interest and appraisals: Multimodal analysis and automatic recognition. Frontiers in ICT **5**, 17 (07 2018). https://doi.org/10.3389/fict.2018.00017
18. Texeira, T., Wedel, M., Pieters, R.: Emotion-induced engagement in internet video advertisements. Journal of Marketing Research **49** (04 2012). https://doi.org/10.2307/23142841

19. Tian, Y.I., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2), 97–115 (2001). https://doi.org/10.1109/34.908962
20. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. Affective Computing, IEEE Transactions on **5**, 86–98 (04 2014). https://doi.org/10.1109/TAFFC.2014.2316163
21. Wu, J., Yang, B., Wang, Y., Hattori, G.: Advanced multi-instance learning method with multi-features engineering and con-servative optimization for engagement intensity prediction (10 2020). https://doi.org/10.1145/3382507.3417959
22. Yang, J., Zhang, F., Chen, B., Khan, S.U.: Facial expression recognition based on facial action unit. In: 2019 Tenth International Green and Sustainable Computing Conference (IGSC). pp. 1–6 (2019). https://doi.org/10.1109/IGSC48788.2019.8957163