Predicting Meeting Success With Nuanced Emotions

Ke Zhou, *Nokia Bell Labs, Cambridge, U.K. and also University of Nottingham, Nottingham, NG7 2RD, U.K.*
Marios Constantinides and Sagar Joglekar, *Nokia Bell Labs, Cambridge, CB3 0FA, U.K.*
Daniele Quercia, *Nokia Bell Labs, Cambridge, U.K. and also King's College, London, WC2R 2LS, U.K.*

While current meeting tools are able to capture key analytics (e.g., transcript and summarization), they do not often capture nuanced emotions (e.g., disappointment and feeling impressed). Given the high number of meetings that were held online during the COVID-19 pandemic, we had an unprecedented opportunity to record extensive meeting data with a newly developed meeting companion application. We analyzed 72 h of conversations from 85 real-world virtual meetings and 256 self-reported meeting success scores. We did so by developing a deep-learning framework that can extract 32 nuanced emotions from meeting transcripts, and by then testing a variety of models predicting meeting success from the extracted emotions. We found that rare emotions (e.g., disappointment and excitement) were generally more predictive of success than more common emotions. This demonstrates the importance of quantifying nuanced emotions to further improve productivity analytics, and, in the long term, employee well-being.

In the workplace, meetings are often considered as one of the primary sources of stress and, as most of the day-to-day communication has shifted online, new stressors are now a reality (e.g., Zoom fatigue). To cope with the sudden transition to online communication, current tools provide informal channels of communication to help participants stay connected,[2] and provide data analytics through audiovisual or textual analyses.[6] However, these tools often fail to capture the nuances of human-to-human communication. Having an unprecedented opportunity to record and obtain meeting conversations, we investigated the extent to which a nuanced emotional classification analysis (compared to the widely adopted sentiment analysis[1,9]) is predictive of meeting's success.

In so doing, we made the following three sets of contributions.

1) We collected 72 h of meeting conversations from 85 real-world, virtual corporate meetings, and 256 self-reported meeting success scores, which we used as the ground truth in our predictive models.
2) We developed metrics that captured 32 nuanced emotions expressed in those meetings[a].
3) We built a model that predicted a meeting's success upon these metrics, and found that rare emotions were more predictive than more common emotions.

At the end of this article, we will discuss the potential uses of these new analytics in current and future tools for monitoring productivity and employee well-being in any organization.

RELATED WORK

Meeting analytics are the key to productivity and health;[7] in a sense, it provides a way to reflect and, ultimately, run meetings more effectively, creating a mentally healthy environment for employees. Previous research on meeting analytics focused on audiovisual analyses; examples include generating speaker-annotated meeting transcripts;[19] identifying dominance and monitoring meeting participants' interactions;[3] and detecting

1536-1268 © 2022 IEEE
Digital Object Identifier 10.1109/MPRV.2022.3145047

[a]Details can be found at: http://social-dynamics.net/FabConvo.

Emotion	Example of the speaker prompt for a conversation associated with the emotion
Afraid	It feels like hitting to blank wall when i see the darkness.
Angry	I lost my job last year and got really mad.
Annoyed	We have a new manager at work and it isn't going well.
Anticipating	Had a interview yesterday. I think it went pretty well. I hope I get the job.
Anxious	Halloween can't get here quick enough.
Apprehensive	My son was invited for a sleepover but we don't know the host family very well.
Ashamed	I never took a drink before and my friends were all over me to take a drink.
Caring	I really enjoyed taking care of my nephew yesterday. His mother had left him with me since she had work.
Confident	A year ago I spent a lot of time doing research for an article. Then a top journal accepted it for publication.
Content	I am ok with being average.
Devastated	My daughters little dog passed away about 4 months ago, it really shook all of us.
Disappointed	A friend lied to me about why he did not come to my moms funeral. It was stupid at the time but I am still upset about it.
Disgusted	I just saw a naked man run through my neighborhood. It was so weird.
Embarrassed	When shopping,I found a pants that I liked, but my partner insisted on getting different one. We argued and people were staring at us.
Excited	I met Elton John at one of his concerts because I brought him flowers to the stage. Not an easy task.
Faithful	i worked for two months and expected nothing, but i wasn't disturbed. I had passion for what i was doing.
Furious	Someone tried to steal my computer a year ago, got super angry with them.
Grateful	My friends set me up a huge surprise party for my birthday.
Guilty	I cheated on my girl friend once and haven't told her.
Hopeful	Slowly but surely the light at the end of tunnel is coming.
Impressed	I was very impressed when my brother came home from college. He was more mature and looked like a grown up.
Jealous	My coworker is allowed to work remotely, but I am not...
Joyful	Planning out my new home has turned out to be a blast!
Lonely	I found myself divorced a few years ago, and for the first time in my life, I was living alone.
Nostalgic	When I hear Christmas music during the holidays I miss my family in the Philippines.
Prepared	I had to board my entire house when the hurricane was coming.
Proud	I'm 26 and can bench, squat, and deadlift over 300lbs...I feel pretty good about myself overall.
Sad	During christmas a few years ago, I did not get any presents.
Sentimental	The first time someone gave me flowers made me feel so special.
Surprised	My friend was supposed to be on her trip for months and months but she came home early! I missed her!
Terrified	The other night I was alone and heard a nose coming from the kitchen... it was creepy.
Trusting	I have had the same best friend for 10 years now.

FIGURE 1. Examples of speaker prompt for each of the 32 emotions in the empathetic dialogues dataset.[16] Each conversation is between two crowd-sourcing platform users and is about a specific situation associated with a specific emotion. Each row corresponds to an emotion class and the corresponding example speaker prompt. These two elements were used for training our emotion classifiers.

action items.[17] Recently, several studies focused on tracking meeting behavior through automated emotional speech classification,[13] group dynamics analysis using speech transcript,[20] and group rapport estimation.[12]

Despite that sentiment analysis and emotion classification have been found to be helpful in inferring the experience (success) of a meeting,[5] most of the previously studied analytics have been based on a limited set of coarse-grained emotions, such as positive or negative sentiment (happy or sad). Yet, as many experienced during the COVID-19 pandemic, meeting analytics would benefit from being able to capture a wider range of emotions. To tackle this challenge, our study quantified 32 emotions[16] during 85 corporate meetings and correlated them with participants' self-reported meeting success.

NUANCED EMOTION DETECTION

Using a crowdsourced dataset of conversational text annotated with 32 emotions, we trained multiple classifiers to extract nuanced emotions from any given text.

Empathetic Dialogues Dataset

To model nuanced emotions within conversations, we leveraged the publicly available empathetic dialogues dataset[16] covering 32 nuanced emotions as opposed to existing approaches (e.g., IEMOCAP, CMU-MOSEI, and RECOLA), which are constrained to a limited number of (typically eight) emotions. The dataset consists of 24,850 conversations, obtained from 810 workers in a crowdsourcing task published on Amazon Mechanical Turk. A pair of workers were asked to: i) select an emotion word each, and describe a situation when they felt that way (we call this description "speaker prompt"); and ii) have a conversation about each of the two situations. The resulting speaker prompts together with their emotion classes were used (see Figure 1) for training our two emotion classifiers. We did not use the entire conversations but just the prompts because the goal of our work was not to build emotion-aware conversational agents (as Rashkin *et al.*[16] aimed to do) but to simply build emotion classifiers.

Classifying Emotions

To identify those 32 emotions in a conversation, we used two classification frameworks: i) a traditional ensemble classifier (AdaBoost), and ii) a deep-learning model [long short-term memory (LSTM)].

Adaptive Boosting (AdaBoost): AdaBoost is an ensemble learning algorithm with gradient boosting.[8]

It uses an iterative approach to learn from the misclassifications of weak classifiers, and it builds a strong classifier by combining multiple weak classifiers; furthermore, it is well-suited to small datasets and makes it easy to interpret the contribution of individual features. To train the AdaBoost model, we extracted four families of paragraph-level textual features. We picked these four families because they have been successfully used to solve a variety of natural language processing (NLP) tasks, are intuitive, and cover several facets of language use. Here, we summarize them shortly and we refer the reader to the original publications[11,14,15] for further details. The first family of features captures aspects of linguistic style: the use of part of speech and many simple syntactic markers.[11] The second one relies on Linguistic Inquiry and Word Count (LIWC),[14] a widely used linguistic lexicon that maps words into linguistic, psychological, and topical categories. The third family of features captures the distribution of vocabularies by counting each sentence's unigrams and bigrams. To reduce the sparsity of the n-gram space, we considered only those that occur 10 times or more in the training set, and we filtered them using log-odd Dirichlet priors to further narrow the set to those n-grams that are highly discriminative. The fourth and final family of features represents each sentence as a 300-dimension GloVe embedding[15] and averages the embeddings of all the words in the sentence. We performed a grid search to tune the learning rate of AdaBoost. In a binary classification task, AdaBoost outputs a [0,1] confidence score that captures the likelihood of the sample belonging to the positive class. Given the nature of our task, we performed the multiclass classification problem by using a one-versus-rest strategy and did so by combining multiple binary classifiers.

LSTM: LSTM[18] is a type of recurrent neural network particularly suited to process data that is structured in temporal or logical sequences. LSTMs have demonstrated to achieve excellent results in time series forecasting as well as in NLP tasks. LSTM accepts fixed-size inputs; in our experiments, we fed one word at a time to it (in the form of a 300-dimension GloVe vector). Each new word updates the model's status by producing a new hidden-state vector. The input sequence is the speaker prompt, and the target value is the prompt's emotion class. Following a standard approach, we applied a linear transformation to reduce the last hidden vector. We experimented with a simple LSTM model with no attention, shortcut connections, or other additions. The architecture had 2 LSTM layers, 256 hidden states, 1 linear layer, and the sigmoid activation layer; the binary cross-entropy was used as the loss function. We performed a grid search to tune its two hyperparameters (i.e., the Adam optimizer's learning rate and the number of epochs). Standard batch normalization was applied to all LSTM layers, and a dropout rate of 0.3 to avoid overfitting was used. Similar to AdaBoost, we deployed the one-versus-rest strategy for combining those binary classifiers for our multiclass classification task.

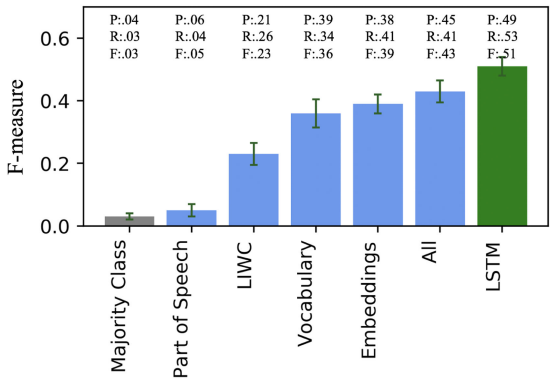

FIGURE 2. Performance in terms of F-measure of different emotion detection models on the empathetic dialogues dataset (weighted average F-measure of all 32 nuanced emotions over 10-fold cross-validation; gray: baseline, blue: AdaBoost, green: LSTM). In addition to the F-measure (F), its composing Precision (P), and Recall (R) are reported on top of each bar.

Experiments

To classify the 32 emotions, we used both AdaBoost and LSTM models. We performed a 10-fold cross-validation, and report each model's performance as the average F-measure of the 10 validation rounds. F-measure is a widely used performance metric as it combines both precision and recall, showing how precise the classifier is and, at the same time, how robust it is.

Results for Emotion Classification

The experimental results are shown in Figure 2. Not surprisingly, the baseline model of simply picking the majority class achieved a subpar performance with a F-measure of merely 0.03. LSTM reached the best performance, yielding top scores on most of the 32 emotions: performances across emotions did vary, ranging from 0.38 to 0.73. Upon performing the 10-fold cross-validations, the resulting F-measure values for the best performing LSTM classifier were: 0.51 (mean), 0.50 (median), and 0.03 (standard deviation). The performance generally dropped for the AdaBoost model, even when relying on all the available features. Among

all the AdaBoost models, the model with part-of-speech features performed the worst, whereas the embedding-based model performed the best. AdaBoost combining all the features achieved the best performance, but only marginally outperformed the embedding- and LIWC-based models.

Although the AdaBoost models yielded competitive performances, they were less effective than the deep-learning LSTM model in tackling the high lexical variety with which certain emotions were expressed. Nevertheless, the nature of the AdaBoost framework allowed us to obtain the importance of its features in predicting emotions, thus providing a human-readable understanding of which parts of the verbal exchanges tended to be predictive. We examined two interpretable AdaBoost models (one based on LIWC and the other based on vocabulary features) and measured feature importance for them. We found that, when examining both models, naturally, sentiment-related LIWC/vocabulary categories were the most important features. These include word categories reflecting anxiety (e.g., scared and haunted), positive emotion (e.g., excited and awesome), perceptual processes (hear and feel), anger, and swearing.

Generally speaking, all the LSTM-based classifiers on the 32 emotions performed quite well. Yet, some performed better than others. To further examine the performance variance across different emotions, we analyzed each emotion classifier (one versus rest), and found that the emotions lonely, jealous, and nostalgic are easier to predict (with F-measure over 0.65), whereas *disappointed*, *angry*, and *ashamed* are comparatively more difficult (with F-measure less than 0.40). We provide the confusion matrix in Figure 3 to broadly illustrate the challenges of classifying the 32 emotions. We found that certain emotions were misclassified into their closest emotion(s), such as *angry* to *furious*, *ashamed* to *guilty*, *apprehensive* to *anxious*, *sad* to *devastated*, and *joyful* to *content* or *excited*. For simplicity, in the rest of this article, we present the results of the best-performing model for classifying emotions—the LSTM framework.

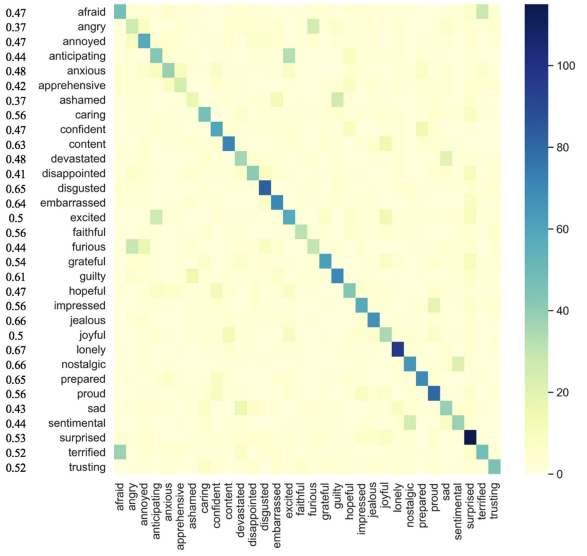

FIGURE 3. Confusion matrix of the actual versus predicted emotions for each of the 32 emotions. The predictions are done by the LSTM model. The elements on the diagonal represent the fraction of posts for which the predicted and the actual emotions were the same (correct classifications), while those off the diagonal represent the fraction of posts for which the "actual" emotion was incorrectly classified to be the "predicted" emotion. Across the rows of the matrix, the prediction performance (F-measure) for each emotion class is shown on the left of the emotion name.

PREDICTING MEETING SUCCESS

After classifying the emotions in our meetings with that framework, we were able to design models that predicted self-reported meeting success scores from the classified emotions.

Meeting Dataset

Using Cisco's WebEx companion platform,[2] we collected data from 85 virtual corporate meetings, approved by the HR department of the authors' institution. In total, these meetings lasted 4373 minutes with a median of 4 people participating in each meeting (min: 2, max: 65, with 11 meetings participated by more than 10 people). The dataset is comprised of a diverse range of meetings with varying duration (min: 20.6 minutes, median: 48.3 minutes, and max: 180.2 minutes), hours of the day (earliest and latest meeting happened at 8 AM and 6 PM, respectively, on that day), days of the week (Mon–Fri), and days of the month (1–31). The companion platform allowed participants to earmark key moments with a mobile app. These moments were then converted into one-minute audio chunks, which the meeting participants could play back in retrospect to get a quick audio summary of the meeting. Earmark moments defined salient moments of the meetings, singling out parts of a meeting its participants considered important.[2]

The companion platform also allowed us to obtain self-reported measures that referred to the participants' meeting experience. More specifically, at the end of each meeting, the participants were prompted to answer two questions: one captured $Q_{\text{psychological}}$, which is the extent to which [a participant] felt

listened or motivated to be involved, and the other captured $Q_{\text{execution}}$, which is the extent to which [a participant] felt that the meeting had a clear purpose and structure. The two questions were answered on a 1–7 Likert scale, with 7 indicating a greater extent. These two questions resulted from an extensive large-scale crowdsourcing study.[4] The goal of that study was to identify the key predictors of a meeting's psychological experience: A 28-item questionnaire was administered to 363 individuals whose answers were then statistically analyzed through principal component analysis (PCA). As a result, the following three main predictors of a meeting's psychological experience were identified: the two previously mentioned (which explained 51% of the variability) plus a third capturing the level of comfort of the physical environment (which explained a further 11% of the variability). Since the present study included only virtual meetings, the third predictor did not apply, while the first two predictors did and, as such, were captured through self-reports. We consequently defined a success score as $\text{success} = (0.759 \cdot Q_{\text{psychological}}) + (0.673 \cdot Q_{\text{execution}})$, where $Q_{\text{psychological}}$ is the average value for the self-reports concerning psychological safety, $Q_{\text{execution}}$ is the average value for the self-reports referring to good execution of the meeting, and the two weights of 0.759 and 0.673 are their loading factors, which resulted from the PCA analysis in Constantinides *et al.*'s work.[4]

The resulting distribution of meeting success had a minimum of 5.5, a median of 7.8, and a maximum of 10.0. Three annotators then performed a manual inspection of the success scores with corresponding meeting recordings and found that the meetings with scores in the [5.5, 7.7] range tended to be indistinguishable in terms of success from each other, and so were the meetings with scores in the [8.0, 10] range. As such, success score could not be taken at face value but had to be binarized; yet, given the two previous ranges, the median value of 7.8 ended up being a natural threshold for the binarization, assigning all meetings to the positive class or negative one (i.e., categorizing them into "successful" and "unsuccessful").

Each meeting in the dataset was then stored as a set of one-minute audio chunks of the earmarked moments plus the participant's self-reported answers and success score. We transcribed the earmarked audio chunks using the state-of-the-art Google's API Speech-to-Text service;[b] each meeting's transcript was used in our textual analyses, as we shall see. In total, all the 85 meetings contained 1007 earmarked moments (a meeting on average had 11 earmarked moments), and 256 answers to the two questions.

[b]Speech-to-Text API: https://cloud.google.com/speech-to-text. It has been found that Google has superior performance on speech recognition compared to other platforms and tools.[10]

Predicting Success

To predict a meeting's binarized success score, we used the meeting's predicted emotions (i.e., its 32-dimensional emotion vector) as input, and tested various classifiers, including logistic regression, a support vector machine, a random forest, a XGBoost, and an AdaBoost classifier. We chose these classifiers as they represent a wide range of well-performing linear and nonlinear classification algorithms. We measured performance using a standard classification metric, that is, the area under the curve (AUC) and employed a leave-one-out cross-validation. We report the averaged AUC, which is the mean value of the AUCs resulting from all the leave-one-out validation rounds. The best-performing model was AdaBoost: for brevity, we report only its results.

Given that we had a relatively small dataset and many nuanced emotion features, we performed two feature selection methods to reduce the dimensionality and to avoid overfitting: ANOVA F-test and recursive feature elimination (RFE). analysis of variance (ANOVA) is a parametric statistical hypothesis test for determining whether the means from two or more samples of data come from the same distribution or not. ANOVA uses F-tests to statistically test the equality of means, determining the independence between numerical features, and categorical target classes and eliminating those. On the other hand, RFE makes feature selection by iteratively training a set of data with the current set of features and eliminating the least significant feature indicated. This procedure was performed solely on the training dataset (fold) on AdaBoost and was repeated until a specified number of features remained. We carried out both feature selection methods, and, among all the five classifiers, we found that the AdaBoost model combining the top 10 emotions selected by RFE performed the best.

For comparison's sake, in addition to that AdaBoost model, we evaluated models that predicted success from the following features alternative to emotions.

1) *Meeting characteristics* (used as control): meeting type (*ad hoc*, recurring, and scheduled), meeting size (number of participants), and meeting duration.
2) *Sentiment* (used as a baseline): both Valence Aware Dictionary and sEntiment Reasoner, a dictionary-based approach[9], and Flair (based on

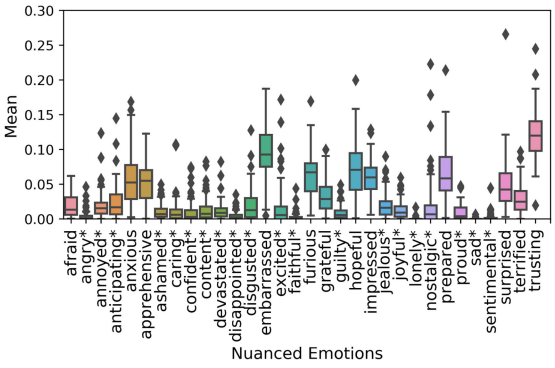

FIGURE 4. Probabilities of the 32 nuanced emotions being expressed in our meetings. Each box plot shows the mean and variance of the probability. The emotions labeled with "*" are uncommon, nuanced emotions with skewed frequency distributions. For example, *embarrassed* and *trusting* are two common emotions, while *disappointed* and *faithful* are two rare emotions, which occurred only in a few meetings.

deep-learning)[1] were applied to meeting transcripts for classifying the sentiment scores.

3) *Text embeddings* (used as a baseline): 300-dimension GloVe embeddings were used to represent a meeting's transcript by averaging the embeddings of all the words that formed the transcript.

Results for Meeting Success Prediction

First, we inspected the distributions of those nuanced emotions across all meetings (see Figure 4). As one expects, some of those emotions were more frequently conveyed by meeting participants, such as trusting, embarrassed, anxious, and hopeful, while other emotions were rarer, such as *angry, sad,* and *guilty*. However, we can observe that the variance of emotions expressed across meetings was relatively high, demonstrating that different meetings were characterized by different emotions.

By then looking at the evaluation results for different models that leveraged meeting characteristics, sentiment, text embeddings, and nuanced emotions for predicting meeting success [see Figure 5(a)], we observed that the model that only incorporated the meeting characteristics (control) performed the worst (AUC = 0.49), whereas the model that leveraged traditional sentiments performed only marginally better (AUC = 0.56). The baseline model that leveraged text embeddings performed better than other baselines (AUC = 0.63). When incorporating all the nuanced emotions, the performance (AUC = 0.62) increased compared to the sentiments-based model, at par with the performance of the text embedding classifier. However, by only considering the top 10 emotions based on RFE feature selection [see Figure 5(a)], the performance was boosted further to an AUC of 0.73. Given the relatively small size of our dataset, it comes as no surprise that feature selection helped reduce our model's overfitting.

As we have seen in the distribution plots of Figure 4, the combinations of emotions significantly varied across meetings and, as such the classifier

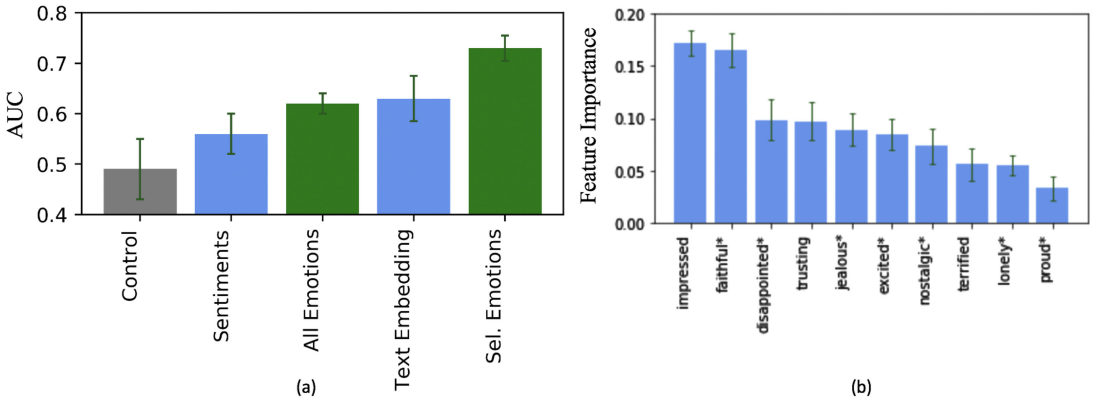

FIGURE 5. Evaluation of meeting success prediction using nuanced emotions: (a) Performance (AUC) of our models trained on meeting characteristics (gray), on sentiment, on text embedding (blue), and on nuanced emotions (green). "All Emotions" refers to the model that leveraged all the nuanced emotions as features, whereas "Sel. Emotions" refers to the model that leveraged the 10 most predictive emotions (based on the RFE feature selection method). (b) Feature importance (absolute) of the AdaBoost model predicting meeting success from a specific emotion. The emotions labeled with "*" had skewed frequency distributions, highlighting that they were relatively uncommon.

Emotion	Examples
Impressed	"Yeah, he's **definitely flattered** because you can see all the US cities are in the horizontal line.", "Interestingly, the ranking is there at least in US, right?" "Yeah. You are right. "
Faithful	"I feel like it's not a bad time to have a reality check to make sure that we're moving in the right direction.", "From everything I've seen and heard, **I'm extremely positive that we, as an organization, we've always prioritized things that are important and we should keep prioritizing them.**"
Disappointed	"**No, all those do not fit there.** I haven't read it and I don't know if I missed anything, but that's my current opinion.", "From the results, we can sort of extract the heart rate, but it doesn't really matter very much.", "**I would expect so** given you have tried so many variables."
Trusting	"Yeah, so the reviews were good. Actually, I think we were **very close to acceptance** as far as I remember." "Yeah. **I trust it is the case.** It says a review score of 5", "Please email the program chair and they would reply hopefully soon."
Jealous	"Democratic Congressman offered a challenge to President Trump. I'll say this to the President Trump.", "You want to investigate Joe Biden, go ahead do it. Do it hard do it dirty.", "**Do it the way you do it.** Just don't do it by asking a foreign leader to help you in your campaign."
Excited	"So are you guys into those songs?", "Not really, but I'm **happy** not too many people are around. That's **one of the best things**. Yeah.", "Yeah, right. Thanks and have **a good one!**"
Nostalgic	"We need to know about the **past experience to help us recreate it**. But it's not just about the one, it is about other live experience in the software. Through that way **we will create the memory and meaning**. So I recommend this to you."
Terrified	"Have you seen the Italian outbreak now?", "No, what's happening now?", "475 **people died** in one day", "**Oh my God**, the death growth is super exponential."
Lonely	"So we still have **other people not connected**. There may be another one more and we are ready.", "So please make sure that your Bluetooth is on because we communicate via Bluetooth communication.", "So **who is missing**, guys?"
Proud	""What do you think about the assignment?","Excellent.","Thank you very much guys for that. I think **we really really got far better compared to before!**"

FIGURE 6. Excerpts of examples taken from our meeting transcripts, which were automatically classified with a specific emotion. Names appearing in the original dialogues were paraphrased, and quotes in boldface indicate language markers likely related to the emotion.

might have well captured nonlinear relationships among these emotions, which could have then contributed to the ultimate prediction accuracy. Therefore, we inspected the feature importance of the best performing AdaBoost model trained on specific emotions [see Figure 5(b)], allowing us to identify which emotions contributed the most to the prediction accuracy. One interesting and surprising finding lies in the discriminative power of the nuanced emotions: we found that, for the best model, out of the ten most predictive emotions, seven were rare, and only three were common [see Figure 5(b)]. This demonstrates that the emotions that occurred very frequently across meetings might not have been as predictive as the combinations of emotions that occurred less frequently yet tended to make each meeting unique (see Figure 6). For example, based on the presence of faithfulness and disappointment, one could easily distinguish a successful meeting from an unsuccessful one. By contrast, based on the presence of more common emotions, such as prepared or hopeful, one would find it hard to make such a distinction.

DISCUSSION AND CONCLUSION

We showed that nuanced emotions expressed in conversations were associated with self-reported meeting success. Compared to coarse-grained sentiment analysis, these emotions could potentially enrich meeting analytics, both in real-time and postmeeting. Interestingly, we found that certain rare emotions, such as *disappointing,* were generally more predictive of meeting success than common emotions, highlighting the importance of quantifying even those emotions that do not frequently occur.

Our work has several implications. First, our quantification of nuanced emotions could be widely adopted in organizational and management research. As these emotions greatly matter in meetings, if captured, they could help teams create psychologically safe environments. Second, our models could be deployed and integrated with any communication tool (e.g., Meetcues[2]) that provides transcripts or voice recordings. More efficient meetings might help reduce employees' cognitive overload and improve their mental health.

This work has five main limitations that call for future research efforts. First, our dataset draws from business meetings; thus, our findings might not generalize to other types of meetings. Future work includes: i) applying our trained model in other types of meetings; and ii) building more nuanced prediction models that move beyond dichotomized meeting success. Second, we adopted meeting transcripts as a basis

upon which to extract analytics. However, other aspects derived from facial expressions or body language could be informative (e.g., key turning points in a meeting). Third, our nuanced emotions are not specific to meetings. For example, in the specific context of meetings, certain emotions (e.g., *jealous*) were likely confused with similar emotions (e.g., *disappointed*). Tailoring emotion classification to the meeting context may well boost performance, pushing our model's fairly high AUC even further up. Fourth, although the meeting earmarked moments enabled us to perform meeting analytics in real-time, these moments captured only specific portions of a meeting, not necessarily being representative of the entire meeting. Finally, given the limited dataset, we could not study whether emotions unfolded during a meeting in predictable ways. Based on manual inspection, we found that highly successful meetings tended to start with excited and to end with impressed, while unsuccessful ones tended to end with *disappointment*. These anecdotal results suggest that the study of the evolution of emotions during a meeting could be a promising research direction.

In the future, by monitoring nuanced emotions in different communication channels within an organization (e.g., company and university), one could track organizational productivity and well-being, and proactively deploy a suite of ameliorative interventions at both the individual and organizational levels. However, despite the benefits such tracking technologies may bring, they have to be thoughtfully deployed, being mindful of ethical considerations.

REFERENCES

1. A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art NLP," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (Demonstrations)*, 2019, pp. 54–59, doi: 10.18653/v1/N19-4010.
2. B. A. Aseniero, M. Constantinides, S. Joglekar, K. Zhou, and D. Quercia, "MeetCues: Supporting online meetings experience," in *Proc. IEEE Visual. Conf.*, 2020, pp. 236–240, doi: 10.1109/VIS47514.2020.00054.
3. C. Busso, P. G. Georgiou, and S. S. Narayanan, "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 2, pp. II-685–II-688, doi: 10.1109/ICASSP.2007.366328.
4. M. Constantinides, S. Šćepanović, D. Quercia, H. Li, U. Sassi, and M. Eggleston, "ComFeel: Productivity is a matter of the senses too," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–21, 2020, doi: 10.1145/3432234.
5. E. Coutinho and N. Dibben, "Psychoacoustic cues to emotion in speech prosody and music," *Cogn. Emotion*, vol. 27, no. 4, pp. 658–684, 2013, doi: 10.1080/02699931.2012.732559.
6. A. J. Cowell *et al.*, "Understanding the dynamics of collaborative multi-party discourse," *Inf. Visual.*, vol. 5, no. 4, pp. 250–259, 2006, doi: 10.1057/palgrave.ivs.9500139.
7. P. Gonzalez-Alonso, R. Vilar, and F. Lupiáñez-Villanueva, "Meeting technology and methodology into health big data analytics scenarios," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst.*, 2017, pp. 284–285, doi: 10.1109/CBMS.2017.71.
8. T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009, doi: 10.4310/SII.2009.v2.n3.a8.
9. C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, 2014, vol. 8, pp. 216–225.
10. V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)," *Int. J. Eng. Res. Appl.*, vol. 7, no. 3, pp. 20–24, 2017, doi: 10.9790/9622-0703022024.
11. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60, doi: 10.3115/v1/P14-5010.
12. P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *Proc. 23rd Int. Conf. Intell. User Interfaces*, 2018, pp. 153–164, doi: 10.1145/3172944.3172969.
13. E. R. O'Neill, M. N. Parke, H. A. Kreft, and A. J. Oxenham, "Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 149, no. 2, pp. 1224–1239, 2021, doi: 10.1121/10.0003532.
14. J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Univ. Texas Austin, Austin, TX, USA, Tech. Rep., 2015.

15. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
16. H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5370–5381, doi: 10.18653/v1/P19-1534.
17. S. Somasundaran, J. Ruppenhofer, and J. Wiebe, "Detecting arguing and sentiment in meetings," in *Proc. 8th SIGdial Workshop Discourse Dialogue*, 2007, pp. 26–34.
18. M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 194–197, doi: 10.1.1.248.4448.
19. T. Yoshioka *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 276–283, doi: 10.1109/ASRU46091.2019.9003827.
20. N. Zhang, T. Zhang, I. Bhattacharya, H. Ji, and R. J. Radke, "Visualizing group dynamics based on multiparty meeting understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 96–101, doi: 10.18653/v1/D18-2017.

KE ZHOU is currently a Senior Research Scientist with Nokia Bell Labs, Cambridge, U.K., and an Assistant Professor of computer science with the University of Nottingham, Nottingham, U.K. His research interests include information retrieval and user modeling. He is the corresponding author of this article. Contact him at ke.zhou@nokia-bell-labs.com.

MARIOS CONSTANTINIDES is currently a Senior Research Scientist with Nokia Bell Labs, Cambridge, U.K. His research interests include human–computer interaction and ubiquitous computing. Contact him at marios.constantinides@nokia-bell-labs.com.

SAGAR JOGLEKAR is currently a Research Scientist with Nokia Bell Labs, Cambridge, U.K. His research focuses on representation learning in the areas of social computing and urban informatics. Contact him at sagar.joglekar@nokia-bell-labs.com.

DANIELE QUERCIA is currently the Department Head with Nokia Bell Labs, Cambridge, U.K., and a Professor of urban informatics with King's College, London, U.K. His research interests include computational social science and urban informatics. Contact him at quercia@cantab.net.