1-7-2021

Automated rating of patient and physician emotion in primary care visits

Jihyun Park

Abhishek Jindal

Patty Kuo

Michael Tanana

Jennifer Elston-Lafata
Henry Ford Health, jlafata1@hfhs.org

See next page for additional authors

Recommended Citation

Park J, Jindal A, Kuo P, Tanana M, Lafata JE, Tai-Seale M, Atkins DC, Imel ZE, and Smyth P. Automated rating of patient and physician emotion in primary care visits. Patient Educ Couns 2021.

Authors

Jihyun Park, Abhishek Jindal, Patty Kuo, Michael Tanana, Jennifer Elston-Lafata, Ming Tai-Seale, David C. Atkins, Zac E. Imel, and Padhraic Smyth

Patient Education and Counseling xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

Automated rating of patient and physician emotion in primary care visits

Jihyun Park[a,b,1,*], Abhishek Jindal[a,c,1], Patty Kuo[d], Michael Tanana[e], Jennifer Elston Lafata[f,g], Ming Tai-Seale[h], David C. Atkins[i], Zac E. Imel[d], Padhraic Smyth[a]

[a] Department of Computer Science, University of California, Irvine, USA
[b] Apple Inc., Cupertino, USA
[c] Hewlett Packard Enterprise, San Jose, USA
[d] Department of Educational Psychology, University of Utah, Salt Lake City, USA
[e] Social Research Institute, University of Utah, Salt Lake City, USA
[f] Division of Pharmaceutical Outcomes and Policy, University of North Carolina at Chapel Hill, USA
[g] Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, USA
[h] Department of Family Medicine and Public Health, University of California, San Diego, USA
[i] Department of Psychiatry and Behavioral Science, University of Washington, Seattle, USA

ARTICLE INFO

Article history:
Received 8 July 2020
Received in revised form 3 January 2021
Accepted 4 January 2021

Keywords:
Doctor-patient communication
Patient-physician communication
Doctor-patient conversation
Machine learning
Natural language processing
Sentiment analysis
Emotion classification
Primary care visit

ABSTRACT

Objective: Train machine learning models that automatically predict emotional valence of patient and physician in primary care visits.
Methods: Using transcripts from 353 primary care office visits with 350 patients and 84 physicians (Cook, 2002 [1], Tai-Seale et al., 2015 [2]), we developed two machine learning models (a recurrent neural network with a hierarchical structure and a logistic regression classifier) to recognize the emotional valence (positive, negative, neutral) (Posner et al., 2005 [3]) of each utterance. We examined the agreement of human-generated ratings of emotional valence with machine learning model ratings of emotion.
Results: The agreement of emotion ratings from the recurrent neural network model with human ratings was comparable to that of human-human inter-rater agreement. The weighted-average of the correlation coefficients for the recurrent neural network model with human raters was 0.60, and the human rater agreement was also 0.60.
Conclusions: The recurrent neural network model predicted the emotional valence of patients and physicians in primary care visits with similar reliability as human raters.
Practice implications: As the first machine learning-based evaluation of emotion recognition in primary care visit conversations, our work provides valuable baselines for future applications that might help monitor patient emotional signals, supporting physicians in empathic communication, or examining the role of emotion in patient-centered care.

1. Introduction

There are over 900 million medical office visits a year [4]. The quality of patient-physician interactions during these visits, and responsiveness to patient concerns, is essential to patient-centered care [5–9]. Patient-physician interactions involve more than just discussions of symptoms and biomedical facts. Patients can be in intense emotional states when visiting their physicians. How physicians recognize and respond to patient emotions [10–14] is related to important patient outcomes [15,16], including satisfaction with medical care [17,18], adherence to treatment [19], malpractice claims [20], quality of life [21], and other clinical outcomes (e.g., diabetes [22], blood pressure [23]). Consequently, an important aspect of patient-centered care is effective recognition of emotions during medical appointments.

Patients desire that physicians devote more time in medical appointments to their emotional concerns [24]. In fact, about 20 percent of primary care appointments have indicators of mental health concerns (e.g. previous diagnoses, psychotropic medication), where patients may be experiencing high levels of psychological distress and emotions that they bring to

* Corresponding author at: 1 Apple Park Way, [924-4IST], Cupertino, CA, 95014, USA.
E-mail addresses: jihyunp@ics.uci.edu (J. Park), zac.imel@utah.edu (Z.E. Imel), smyth@ics.uci.edu (P. Smyth).
[1] Most of the work was done while the authors were students at the University of California, Irvine.

https://doi.org/10.1016/j.pec.2021.01.004
0738-3991/

appointments [25–27]. Furthermore, patients in appointments may also be experiencing emotions in appointments directly connected to their physical ailments. However, physicians are tasked with responding to a variety of demands during visits, and may miss key emotional moments that in turn impact patient outcomes (e.g. treatment satisfaction, adherence) [28–30]. Tools that can quickly assess patient emotions during medical appointments may help providers better attend to emotions that arise. At present, there are currently limited methods of measuring patient emotions during medical appointments. Patient emotions in medical settings are primarily evaluated using surveys that measure general experiences in treatment [31], which place a burden on the patient, are subject to social desirability and recall biases [32–34]. Direct observation and feedback is the gold standard for supporting physicians in attuning to emotions during medical appointments, but it is time-consuming, expensive, and generally not feasible in most clinical settings [35].

Quantitative text analysis methods can facilitate more direct assessment of emotional events in primary care, leveraging recent advances [36] in natural language processing (NLP) and machine learning to analyze transcripts of patient-physician conversations. Developing NLP models that identify patient expressions of emotion and their valence from visits could contribute to the development of scalable quality improvement efforts by supporting communication that is more informed by emotional attunement to patients, or highlighting important moments that were not addressed. NLP techniques have been used to extract treatment information from electronic health records (EHRs) [37–39], identify post-operative complications [38], as well as identify patients that may be in need for further treatment or care [39–41].

	Utterance	Avg	Emotional Valence by Human Raters					
			4	5	9	10	13	14
0	MD: Why hello there.	1.3	2	1	1	1	2	1
1	How are you doing?	1.0	2	1	1	1	0	1
2	PT: Doing good.	1.5	2	1	2	1	2	1
3	MD: Are you?	0.3	1	0	0	0	0	1
4	PT: Doing good, doing good.	1.2	1	1	1	1	2	1
5	Just, you know, got a little more pressure on me.	-0.8	-1	-1	-1	-1	-1	0
6	MD: Yeah.	0.0	0	0	0	0	0	0
7	PT: My stepfather passed June 5th.	-2.0	-2	-1	-2	-1	-3	-3
8	MD: Oh, I'm sorry to hear that.	-1.3	-1	-1	-1	-1	-2	-2
9	PT: So, I'm running, like twice a week I'm going to my mothers out there by Metro, airport, so.	-0.7	-1	0	-1	0	-2	0
10	She's hanging in there, she was married 40 years.	-1.0	-1	-1	-1	-1	-2	0
11	MD: Okay.	-0.2	0	0	0	0	-1	0
	…							
48	MD: So you 're trying to help your mom out.	-0.3	-1	0	0	0	-1	0
49	PT: Right, right.	-0.3	-1	0	0	0	-1	0
50	MD: Yeah.	-0.2	0	0	0	0	-1	0
51	PT: Yeah.	-0.2	0	0	0	0	-1	0
52	MD: Well it seems like your, your blood pressure looks pretty good, yeah?	1.3	1	1	1	2	2	1
53	PT: I'm hanging in there, still roller skating, still swimming, and still walking.	1.3	1	1	1	2	2	1
54	MD: Are you?	1.0	1	0	0	2	2	1
55	You 're pretty active.	1.7	1	1	1	3	2	2
56	PT: Trying to be.	0.8	1	1	0	1	1	1
57	MD: That's darn good.	2.2	2	2	2	2	3	2
58	PT: Trying to be.	1.2	1	1	1	2	1	1
	…							
68	MD: When you check your blood sugar, how often do you check it?	0.0	0	0	0	0	0	0
69	PT: Twice a day, in the morning and in the evening.	0.0	0	0	0	0	0	0
70	MD: Okay.	0.0	0	0	0	0	0	0
71	In the morning what kind of range do you get with the numbers?	0.0	0	0	0	0	0	0
72	PT: It ranges between 90 and 110.	0.0	0	0	0	0	0	0
73	MD: Oh that's good.	1.3	2	1	1	1	2	1
74	That's quite good.	2.0	2	1	2	2	3	2

Fig. 1. An example of a section of dialog from a particular visit. The six columns on the right side show the emotional valence ratings assigned by 6 raters. The averaged rating is shown in the third column.

Recently, NLP techniques have been applied to transcripts of medical appointments to identify topics of conversation in settings such as pediatrics [42], HIV care [43,44], psychotherapy [45], and primary care [46,47]. There is recent work classifying emotions using lexical features in married couples' discussions [48]. Other work has focused on classifying emotional valence for in-person [49], and online text-based psychotherapy for depression [50]. However, there has been relatively little work on development and evaluation of NLP technology to automatically capture the emotional content of a medical visit.

There has been a significant amount of work in computer science on emotion and sentiment detection in written text, e.g., product reviews [51] and social media posts [52]. However, this work has limited applicability to primary care conversations. Of more relevance is recent work in NLP on classifying the emotional state of a speaker for each utterance, using lexical features (i.e., words and short phrases) extracted from transcripts of human-human or human-computer dialog [53–55]. This work on classifying speaker state, however, has tended to focus on datasets consisting of relatively short exchanges (e.g., AVEC [56], MELD [57], IEMOCAP [57,58]) often involving simulated conversation topics which are quite different in nature to real-world clinical dialog.

In the current study, we evaluated the performance of machine learning models to rate the emotional valence (positive, neutral, negative) [3] of patients and physicians at the utterance level in 353 transcripts of primary care medical visits. This is the first large-scale study of automated emotion recognition from transcripts of primary care conversations. Recent work in automated annotation of utterances within conversations [47,53,57,58] has shown that contextual or sequential information across utterances is crucial in utterance-level analysis, particularly for short utterances. Thus, in our work, we focus on current state-of-the-art sequential models for predicting emotional valence. In particular, we focus on hierarchical recurrent neural networks [47,59] and compare these models to a simpler non-sequential baseline (logistic regression) as well as to human raters.

2. Methods

2.1. Dataset

The dialog transcript dataset used in this paper is a combination of transcripts from two previous studies: The Mental Health Discussion study by Tai-Seale et al. [2,60,61] and the Assessment of Doctor-Elderly Patient Transactions (ADEPT) study by Teresi et al. [1,62]. The dataset consists of 353 human-generated transcripts with 210k utterances from elderly patients' doctor visits. The majority of the patients in the dataset were white (66.2 %) and female (65.6 %), with an average age of 62 years. An utterance is defined as a discrete sentence unit within a talk-turn, that is separated by periods, exclamation marks, and question marks. The rating team was composed of fourteen raters who were affiliated with the University of Utah; three raters were doctoral students, three were post-baccs, and six were undergraduate students. Doctoral students rated utterances as part of their funding, and post-bacc and undergraduates were compensated for their time coding. Raters were asked to rate utterances based on how they thought the speaker was feeling when they spoke the utterance. Raters were able to use preceding utterances as context for how they rated each utterance. Each utterance was rated by 2.3 different raters on average, where the rating takes an integer value that ranges from -3 (very negative) to +3 (very positive), with neutral at 0. Of the fourteen raters, we removed the ratings from four raters whose distributions of assigned ratings were significantly different from the other raters, which accounted for 24.6 % of all the ratings (Supplement A.2 has details). For the ratings assigned by the rest of the raters, 79.2 percent of all the raters were neutral (0). To estimate interrater reliability, the Intraclass Correlation Coefficient (ICC) was found to be 0.90 using a two-way random effects model ICC(3,k).

Fig. 1 shows an example of the ratings for one visit. In this example, some of the talk-turns are split into multiple utterances, and each utterance has multiple emotional valence ratings assigned here by 6 raters. The third column in Fig. 1 shows the mean emotional valence ratings after taking the average of the raters' valence ratings for each utterance. We regard these averaged ratings across the multiple ratings of each utterance as the reference standard, and use them for training and evaluation of our predictive models.

Fig. 2 plots mean emotional valence, averaged over human raters, for each speaker, for the same visit shown in Fig. 1. The x-axis corresponds to utterances, the same value as the first column in Fig. 1. We see that the patient and physician each transitions through different emotional states during the visit, with the patient having more pronounced negative emotion than the physician. As an example, utterances 7, 52, and 57 in Fig. 1 are marked as solid circles in Fig. 2. When the patient talks about his/her recent bereavement (utterance 7), the emotional valence drops

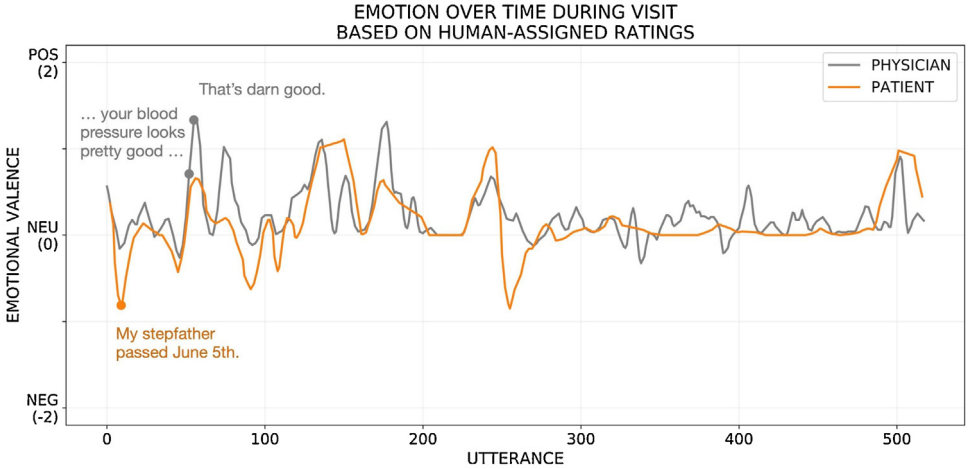

Fig. 2. Per-utterance rater mean emotional valence for a physician (gray) and a patient (orange) during the same visit as in Fig. 1. The mean emotional ratings for each speaker are smoothed with a triangular moving average window size of 7 utterances for visualization purposes. For this visit, 63.7 % of the utterances (330 out of 518 utterances) are from the physician.

to a low (negative value). The physician listens and shows support. After that, the physician tries to brighten the mood by talking about the blood pressure (utterance 52), and compliments the patient for being active and exercising often (utterances 54–57). Similar patterns show up later in this visit as well in other visits.

2.2. Models

Recurrent neural networks [63] are known to be effective in learning from sequential inputs, particularly for language data. They can be used to learn a vector representation of a sentence or an utterance, while keeping the sequential properties of the inputs [64]. We investigated a hierarchical form of a recurrent neural network model that has been used successfully for other tasks in dialog modeling (e.g., Serban et al. [59], Park et al. [47]). The hierarchical nature of the model involves using recurrent networks to model word sequence within an utterance and to model sequences of utterances within a visit. In other words, the model encodes an utterance using the sequence of words in an utterance, while also taking into account the whole sequence of utterances in a visit, including the neighboring utterances, to predict an emotional valence of an utterance. This is particularly useful for short utterances that often contain very little contextual information. The model also includes a "gated recurrent unit," [65] which is a particular type of recurrent network structure.

The parameters (weights) in the model were optimized by minimizing the log-loss (cross-entropy) objective function on the training data using gradient-based search in an end-to-end fashion. The model was trained to predict one of three mutually-exclusive categories: negative, neutral, or positive. Negative (positive) corresponds to average emotion ratings (per utterance across raters) less than -0.5 (greater than +0.5), with neutral corresponding to ratings between -0.5 and +0.5. The model also takes into account who the speaker is, allowing for a speaker effect in the predictions of emotion.

We used logistic regression (LR) classifier as a baseline model for comparison. The logistic regression classifier uses bag-of-words to encode an utterance representation, which does not incorporate any sequential information between words or utterances. Additional details on models and training procedures are provided in Supplement B.

2.3. Evaluation

We evaluated predictive performance using both correlation and ranking measures. For correlation, we computed the Pearson correlation coefficient across utterances between (i) the average human valence rating per utterance and (ii) a model's valence rating per utterance. A model's valence rating was defined as $P(positive \mid utterance) - P(negative \mid utterance)$ where the conditional probabilities were generated by the outputs of the classification model (Hierarchical recurrent neural network and LR).

For ranking-based evaluation, we used R-precision [66] to measure how often the model's most confident predictions match the rankings from human raters. An R-precision score of 1 means

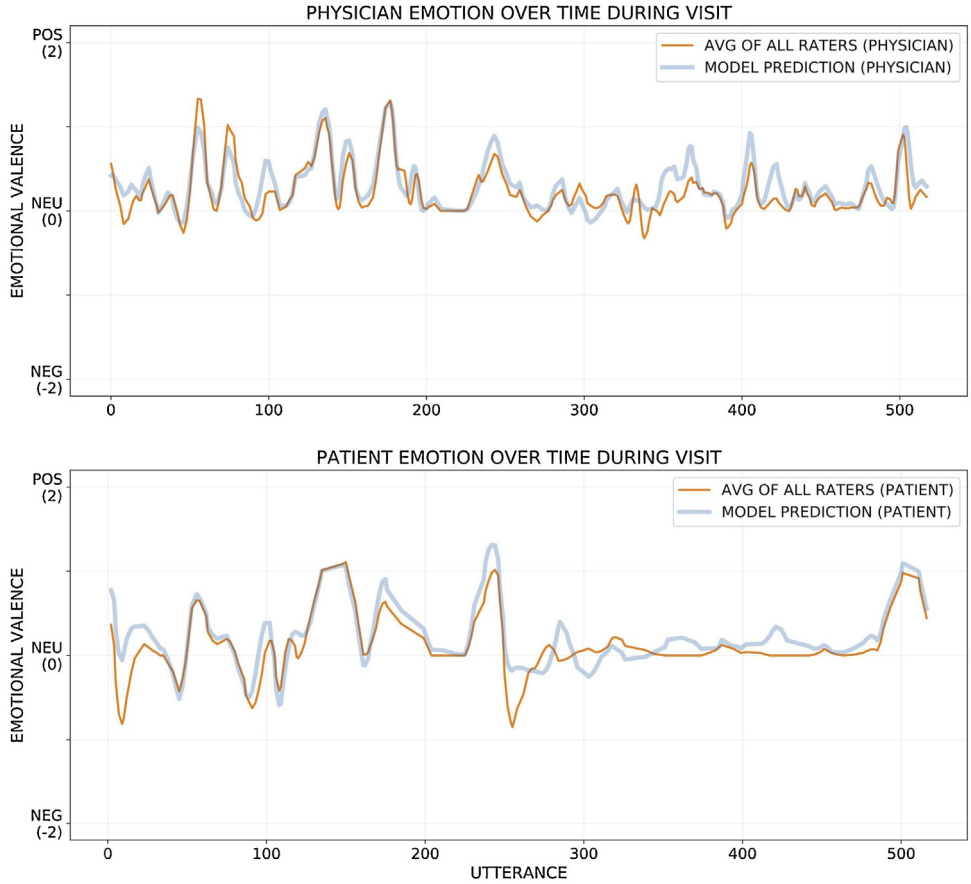

Fig. 3. Examples of emotional valence trajectories of the physician (top) and the patient (bottom) in a visit (the same visit as shown in Fig. 2). Blue lines in both of the plots show the predicted ratings from the model. These ratings were scaled to match the variance of the average human ratings for clarity. Smoothing with a moving average triangular window of size 7 utterances was applied to all of the sequences for visualization purposes (but not used in computing any of the evaluation metrics). The correlation coefficient was 0.831 for the physician (top panel) and was 0.794 for the patient (bottom panel). See Supplement E for more examples.

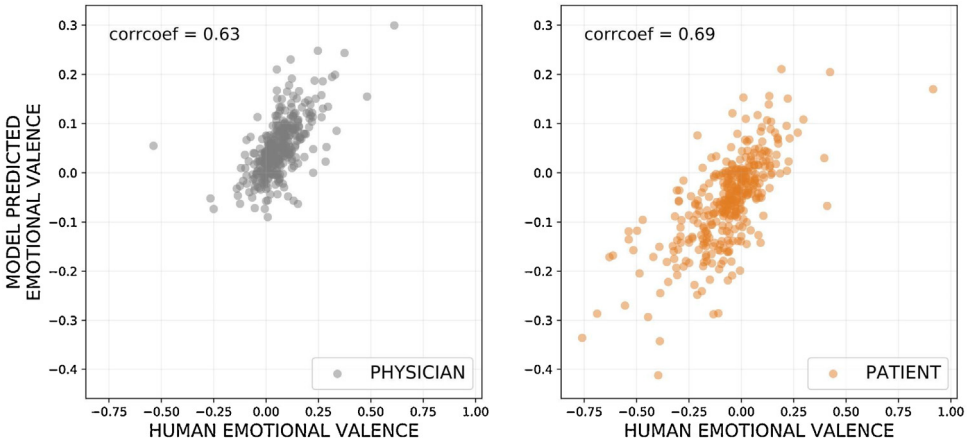

Fig. 4. Mean emotional valence score per visit, for the model (y-axis) versus aggregate human ratings (x-axis), for physicians (left) and patients (right), with Pearson correlation coefficients shown in the top left for each plot.

Table 1
Five example utterances with the highest output probabilities generated from the recurrent neural network model. The output probability from the model and the averaged value of human-assigned ratings for that utterance are shown.

Negative			Positive		
$p(y_{i,j} = neg)$	Human	Text	$p(y_{i,j} = pos)$	Human	Text
0.987	−1.5	To wake up to that woman screaming, terrible, terrible.	0.994	2.0	He's wonderful.
0.978	−2.5	And a lot of times when I'm cleaning houses I'll have such horrible hot flashes.	0.994	2.0	Very good, very good.
0.978	−2.5	And they've been getting worse and worse and worse.	0.993	2.0	That sounds wonderful.
0.975	−2.0	The, oh I felt horrible.	0.992	2.2	Were very happy.
0.975	−3.0	And I hate that, I just hate that.	0.992	2.0	Excellent, excellent.

that the top-ranked R utterances for positive (or negative) emotion, as identified by the model (based on its probabilities), were all rated as positive (or negative) by the human raters; and an R-precision score of 0 means that none of the model's top R utterances were scored as positive (or negative) by the human raters. Additional details on R-precision can be found in Supplement D.2.

In addition to comparing the model predictions with the average of human rater scores, we also computed correlation and R-precision where we compared each human rater's scores to the average of all of the other human rater scores for that utterance, and repeated this for each of the 10 raters, resulting in 10 sets of correlations and R-precisions (see Supplement C for the rater subset information). We refer to these results as OvR (One versus Rest) below.

To obtain the prediction results for all 353 visits, 10-fold cross-validation was used for the two models (the recurrent neural network and logistic regression). The correlation and R-precision values for each of the two models and for the human OvR are computed using the same subsets shown in Supplement D. The numbers for the subsets were reduced to a single number by taking a weighted average of the individual rater OvR numbers, with weights in proportion to the number of utterances rated by each rater.

3. Results

The resulting Pearson correlation coefficients were 0.60 for human OvR, 0.60 for the recurrent neural network, and 0.55 for logistic regression. The averaged R-precisions for the positive class were 0.47, 0.58, and 0.53, for human OvR, recurrent neural network, and logistic regression, respectively. For the negative class, the R-precision scores were 0.44, 0.45, and 0.42, respectively. Results before averaging can be found in Supplement D.

The correlation and R-precision measures for the recurrent neural network model are consistently better than the non-sequential logistic regression baseline model. In addition, for both correlation and R-precision, the recurrent neural network model is comparable with human performance as measured by the human OvR scores.

Fig. 3 shows an example of emotional valence ratings across a visit in one of the test data subsets, for both the predictions of the recurrent neural network model (in blue) and the average of the six human ratings (in orange) for this visit. The two subplots correspond to utterances from the physician (top) and the patient (lower). It is clear for this example in Fig. 3 that the model's predictions track well with the trajectory of the mean human ratings.

3.1. Mean emotional valence per visit

Fig. 4 compares the recurrent neural network model's predictions of emotional valence (y-axis) and the human ratings (x-axis), at the visit level, for both the physician (left) and patient (right). Each datapoint in the plot corresponds to one of 353 visits, where the x and y values are computed by averaging the ratings over utterances in the visit, for the model and for the average of human ratings. There is a strong linear relationship between the emotional valence predicted by the model and the human ratings, at the visit level, for each speaker, with correlation coefficients of 0.63 (physician) and 0.69 (patient). Patient emotion also tends to be more negative than that of physicians, explaining why including

speaker-dependence in a model leads to better predictive performance than a model without speaker dependence (see Supplement F).

3.2. Highly ranked utterances from the model

As a qualitative analysis, Table 1 shows the top five utterances that received the highest negative class probability and highest positive class probability by the model. The results suggest that the model has learned meaningful representation of emotional valence information. The negative examples in Table 1 have very negative expressions and words such as "terrible," "horrible," or "hate." On the other hand, the positive examples include favorable words and expressions such as "wonderful" or "very good." The utterances that returned the highest neutral probabilities were usually related to physical examinations or prescription, and can be seen in Supplement G.

4. Discussion and conclusion

4.1. Discussion

In this work, we used two machine learning models, a recurrent neural network model that takes into account sequential information across utterances in the transcript and a baseline logistic regression model that does not consider sequential information, to automatically rate the emotional valence of patient and physician utterances in transcripts of elderly patients' primary care medical visits. We not only evaluated the performance of these two models by treating the average human rating scores as "true" values but also compared their predictions to the ratings by human raters using one-versus-all approach. Consistent with previous research on the importance of contextual information [47,53], our recurrent neural network model consistently yielded more accurate emotional valence ratings than the logistic regression model. The neural network model predicted patient and physician emotional valence at similar performance levels as human raters, measured by the Pearson correlation coefficient and R-precision.

Although our methods showed promising results, there are limitations mainly due to the data availability. First, our model is only trained using the transcripts from elderly patients where the majority of them need mental health care. More extensive model training and validation with datasets containing a variety of populations who differ in the forms of expressing emotion could be needed. Second, predicting emotional valence is a challenging task since the ratings are highly subjective and hard to be quantified to have a true gold standard. We treated the average ratings from human raters as a reference standard. However, more efforts on creating higher quality labels, such as communicating between the raters to reach a consensus or asking what the speakers were actually feeling in each moment, would be more ideal. Lastly, we were only able to use transcribed texts without nonverbal cues, which are important factors in affective communication [67]. Obtaining a large-scale multi-modal dataset in clinical settings to train a machine learning model is particularly difficult due to privacy and ethical reasons [68], and we leave this as future work.

Machine learning models, that can predict emotional valence of physicians and patients in significantly less time than human raters, have the potential to be applied in medical settings to support larger-scale quality improvement efforts focused on empathic patient-physician communication as well as to scale research on emotional patterns in medical visits. For instance, we could examine how patterns in emotional valence of physicians and patients over the course of appointments are related to important clinical outcomes. Mixed methods research could also be applied to examine how shifts in emotional valence are associated with qualitative differences in physician behaviors. Also, flagging the cases when there is a mismatch in emotional valence of patient and physician could allow researchers to have a deeper analysis on why it happened. Another area where our approach can be applied is the human-machine dialogs, where we assess the quality of the automated responses or regulate to have a desirable emotion.

4.2. Conclusion

In this article, we established baselines for automatically predicting the emotional expression of physicians and patients during primary care visits. Using more than 200k rated utterances, we trained and evaluated the models by exploiting the two different types of outputs: emotional valence scores from the continuous output probabilities and the predicted categorical values.

The performance of the neural network model surpassed a simple logistic regression baseline and approached human-to-human agreement. As machine learning models and automated speech recognition techniques continue to improve, we can expect corresponding improvements in the effectiveness of automated algorithms for characterizing emotions in patient-physician conversations.

4.3. Practice implications

Better understanding the patterns of emotional expressions through which expression of emotions unfolds in appointments can facilitate training of physicians, and improve quality of care. These models could be particularly powerful tools in busy clinical settings where doctors often have large caseloads with limited time [69,70]. Paired with advances in speech recognition [71], machine learning models could facilitate compassionate engagement in patient-centered care by quickly assessing emotional expressions in doctor appointments, and providing real-time feedback to physicians regarding empathy and attunement to patient needs. Future research could focus on developing and integrating machine-learning-based tools into medical settings, and examining the effect of this feedback on physician behaviors.

Author contributions

All authors assisted in the writing of the manuscript. JP conducted the analyses, and was responsible for writing the Methods, Models, and Results sections, with assistance from AJ, PS, and ZI. PS supervised the analyses. MT, PK, and JP were responsible for data management and cleaning. JEL and MT-S were responsible for collection of source data and description of data. PK was responsible for assembling and coordinating the emotion coding team. ZI, MT-S, and PS jointly conceived of the project, and ZEI, PS, and PK wrote the introduction and discussion. DCA provided feedback and consultation regarding methodology and implications of the study.

Funding/support

This work was supported by Patient-Centered Outcomes Research Institute, USA (grant number: ME-1602-34167). Transcripts used in this work were collected under NCI R01-CA112379 and NIA R44-AG015737.

Role of the funder/sponsor

The Patient-Centered Outcomes Research Institute (PCORI) played no part in the design or implementation of the study,

analyses or interpretation of the data, or preparation, review, or submission of the manuscript for submission.

CRediT authorship contribution statement

Jihyun Park: Methodology, Software, Investigation, Visualization, Writing - original draft. **Abhishek Jindal:** Methodology, Software. **Patty Kuo:** Data curation, Writing - original draft. **Michael Tanana:** Data curation, Writing - review & editing. **Jennifer Elston Lafata:** Data curation, Writing - review & editing. **Ming Tai-Seale:** Conceptualization, Resources, Supervision, Writing - review & editing, Funding acquisition. **David C. Atkins:** Writing - review & editing. **Zac E. Imel:** Supervision, Writing - original draft, Project administration, Funding acquisition. **Padhraic Smyth:** Supervision, Writing - original draft, Conceptualization, Methodology.

Declaration of Competing Interest

ZEI, DCA, and MT are co-founders and minority equity stakeholders of a technology company (Lyssn.io) that is focused on developing computational models that quantify aspects of patient-provider interactions in psychotherapy.

Acknowledgments

We gratefully acknowledge Mary Ann Cook for helpful comments and feedback, and Kritzia Merced, Keith Gunnerson, Taylor Shuman, Jimena Murillo, Garrett Battaglia, Grace Thatcher Gardiner, Mateo Connor Fregoso, Azarin Shoghi, Dillon Ely, Jay Junun Park, Joshua Whisenant, Brian Pace, and Paula Gomez, for assistance in emotion coding.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.pec.2021.01.004.

References

[1] M.A. Cook, Assessment of Doctor-Elderly Patient Encounters, Grant No, R44 AG5737-S2, (2002) .
[2] M. Tai-Seale, P.K. Foo, C.D. Stults, Patients with mental health needs are engaged in asking questions, but physicians' responses vary, Health Aff. 32 (2013) 259–267.
[3] J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev. Psychopathol. 17 (2005) 715–734.
[4] NAMCS/NHAMCS - Ambulatory Health Care Data Homepage, (2019) . (Accessed 17 October 2019) https://www.cdc.gov/nchs/ahcd/index.htm.
[5] W. Levinson, Patient-centred communication: a sophisticated procedure, BMJ Qual. Saf. 20 (2011) 823–825.
[6] R.M. Epstein, K. Fiscella, C.S. Lesser, K.C. Stange, Why the nation needs a policy push on patient-centered health care, Health Aff. 29 (2010) 1489–1495.
[7] W. Levinson, C.S. Lesser, R.M. Epstein, Developing physician communication skills for patient-centered care, Health Aff. 29 (2010) 1310–1318.
[8] W. Levinson, K.G. Shojania, Bad experiences in the hospital: the stories keep coming, BMJ Qual. Saf. 20 (2011) 911–913.
[9] K.M. Mazor, B. Gaglio, L. Nekhlyudov, G.L. Alexander, A. Stark, M.C. Hornbrook, K. Walsh, J. Boggs, C.A. Lemay, C. Firneno, C. Biggins, M.A. Blosky, N.K. Arora, Assessing patient-centered communication in cancer care: stakeholder perspectives, J. Oncol. Pract. 9 (2013) e186–93.
[10] C.D. Stults, J. Elston Lafata, L. Diamond, L. MacLean, A.L. Stone, T. Wunderlich, R.M. Frankel, M. Tai-Seale, How do primary care physicians respond when patients cry during routine ambulatory visits? J. Commun. Healthc. 7 (2014) 17–24.
[11] D.K. Chan, T.H. Gallagher, R. Reznick, W. Levinson, How surgeons disclose medical errors to patients: a study using standardized patients, Surgery 138 (2005) 851–858.
[12] E.H.B. Lin, W. Katon, M. Von Korff, T. Bush, P. Lipscomb, J. Russo, E. Wagner, Frustrating patients, J. Gen. Intern. Med. 6 (1991) 241–246.
[13] W. Levinson, W.B. Stiles, T.S. Inui, R. Engle, Physician frustration in communicating with patients, Med. Care 31 (1993) 285–295.
[14] L.M.L. Ong, M.R.M. Visser, F.B. Lammes, J.C. De Haes, Doctor–patient communication and cancer patients' quality of life and satisfaction, Patient Educ. Couns. 41 (2000) 145–156.
[15] R.L. Street, G. Makoul, N.K. Arora, R.M. Epstein, How does communication heal? Pathways linking clinician–patient communication to health outcomes, Patient Educ. Couns. 74 (2009) 295–301.
[16] M.A. Stewart, Effective physician-patient communication and health outcomes: a review, CMAJ 152 (1995) 1423–1433.
[17] W. Levinson, R.M. Frankel, D. Roter, M. Drum, How much do surgeons like their patients? Patient Educ. Couns. 61 (2006) 429–434.
[18] H.-C. Weng, J.F. Steed, S.-W. Yu, Y.-T. Liu, C.-C. Hsu, T.-J. Yu, W. Chen, The effect of surgeon empathy and emotional intelligence on patient satisfaction, Adv. Health Sci. Educ. Theory Pract. 16 (2011) 591–600.
[19] J.N. Fuertes, A. Mislowack, J. Bennett, L. Paul, T.C. Gilbert, G. Fontan, L.S. Boylan, The physician–patient working alliance, Patient Educ. Couns. 66 (2007) 29–36.
[20] W. Levinson, D.L. Roter, J.P. Mullooly, V.T. Dull, R.M. Frankel, Physician-patient communication. The relationship with malpractice claims among primary care physicians and surgeons, JAMA 277 (1997) 553–559.
[21] J. Oates, W.W. Weston, J. Jordan, The impact of patient-centered care on outcomes, Fam. Pract. 49 (2000) 796–804.
[22] S. Nam, C. Chesla, N.A. Stotts, L. Kroon, S.L. Janson, Barriers to diabetes management: patient and provider factors, Diabetes Res. Clin. Pract. 93 (2011) 1–9.
[23] J.E. Orth, W.B. Stiles, L. Scherwitz, D. Hennrikus, C. Vallbona, Patient exposition and provider explanation in routine interviews and hypertensive patients' blood pressure control, Health Psychol. 6 (1987) 29–42.
[24] B. Löwe, U. Schulz, K. Gräfe, S. Wilke, Medical patients' attitudes toward emotional problems and their treatment. What do they really want? J. Gen. Intern. Med. 21 (2006) 39–45.
[25] J.S. Harman, P.J. Veazie, J.M. Lyness, Primary care physician office visits for depression by older Americans, J. Gen. Intern. Med. 21 (2006) 926–930.
[26] R. Mojtabai, Unmet need for treatment of major depression in the United States, Psychiatr. Serv. 60 (2009) 297–305.
[27] QuickStats: Percentage. of Mental Health–Related* Primary Care† Office Visits, by Age Group — National Ambulatory Medical Care Survey, United States, 2010, (2014) . (Accessed 26 September 2019) https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6347a6.htm.
[28] J.A. Hall, D.L. Roter, D.C. Blanch, R.M. Frankel, Nonverbal sensitivity in medical students: implications for clinical interactions, J. Gen. Intern. Med. 24 (2009) 1217–1222.
[29] P.N. Butow, R.F. Brown, S. Cogar, M.H.N. Tattersall, S.M. Dunn, Oncologists' reactions to cancer patients' verbal cues, Psychooncology 11 (2002) 47–58.
[30] C.L. Bylund, G. Makoul, Examining empathy in medical encounters: an observational study using the empathic communication coding system, Health Commun. 18 (2005) 123–140.
[31] M.K. Venetis, J.D. Robinson, K.L. Turkiewicz, M. Allen, An evidence base for patient-centered cancer care: a meta-analysis of studies of observed communication between cancer specialists and their patients, Patient Educ. Couns. 77 (2009) 379–383.
[32] T.E. Burroughs, B.M. Waterman, D. Gilin, D. Adams, J. McCollegan, J. Cira, Do on-site patient satisfaction surveys bias results? Comm. J. Qual. Patient Saf. 31 (2005) 158–166.
[33] T. Heidegger, D. Saal, M. Nuebling, Patient satisfaction with anaesthesia care: what is patient satisfaction, how should it be measured, and what is the evidence for assuring high patient satisfaction? Best Pract. Res. Clin. Anaesthesiol. 20 (2006) 331–346.
[34] A. Gayet-Ageron, T. Agoritsas, L. Schiesari, V. Kolly, T.V. Perneger, Barriers to participation in a patient satisfaction survey: who are we missing? PLoS One 6 (2011)e26852.
[35] Z.E. Imel, M. Steyvers, D.C. Atkins, Computational psychotherapy research: scaling up the evaluation of patient-provider interactions, Psychotherapy 52 (2015) 19–30.
[36] J. Hirschberg, C.D. Manning, Advances in natural language processing, Science 349 (2015) 261–266.
[37] D.S. Carrell, R.E. Schoen, D.A. Leffler, M. Morris, S. Rose, A. Baer, S.D. Crockett, R. A. Gourevitch, K.M. Dean, A. Mehrotra, Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings, J. Am. Med. Inform. Assoc. 24 (2017) 986–991.
[38] H.J. Murff, F. FitzHenry, M.E. Matheny, N. Gentry, K.L. Kotter, K. Crimin, R.S. Dittus, A.K. Rosen, P.L. Elkin, S.H. Brown, T. Speroff, Automated identification of postoperative complications within an electronic medical record using natural language processing, JAMA 306 (2011) 848–855.
[39] S.V. Wang, J.R. Rogers, Y. Jin, D.W. Bates, M.A. Fischer, Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention, J. Am. Med. Inform. Assoc. 24 (2017) 339–344.
[40] P.L. Teixeira, W.-Q. Wei, R.M. Cronin, H. Mo, J.P. VanHouten, R.J. Carroll, E. LaRose, L.A. Bastarache, S.T. Rosenbloom, T.L. Edwards, D.M. Roden, T.A. Lasko, R.A. Dart, A.M. Nikolai, P.L. Peissig, J.C. Denny, Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals, J. Am. Med. Inform. Assoc. 24 (2017) 162–171.
[41] J.F. Ludvigsson, J. Pathak, S. Murphy, M. Durski, P.S. Kirsch, C.G. Chute, E. Ryu, J. A. Murray, Use of computerized algorithm to identify individuals in need of testing for celiac disease, J. Am. Med. Inform. Assoc. 20 (2013) e306–10.
[42] A. Kotov, M. Hasan, A. Carcone, M. Dong, S. Naar-King, K. BroganHartlieb, Interpretable probabilistic latent variable models for automatic annotation of clinical text, AMIA Annu. Symp. Proc. 2015 (2015) 785–794.
[43] B.C. Wallace, M.B. Laws, K. Small, I.B. Wilson, T.A. Trikalinos, Automatically annotating topics in transcripts of patient-provider interactions via machine learning, Med. Decis. Making 34 (2014) 503–512.

[44] E. Mayfield, M.B. Laws, I.B. Wilson, C. Penstein Rosé, Automating annotation of information-giving for analysis of clinical conversation, J. Am. Med. Inform. Assoc. 21 (2014) e122–8.
[45] G. Gaut, M. Steyvers, Z.E. Imel, D.C. Atkins, P. Smyth, Content coding of psychotherapy transcripts using labeled topic models, IEEE J. Biomed. Health Inform. 21 (2017) 476–487.
[46] A. Rajkomar, A. Kannan, K. Chen, L. Vardoulakis, K. Chou, C. Cui, J. Dean, Automatically charting symptoms from patient-physician conversations using machine learning, JAMA Intern. Med. 179 (2019) 836–838.
[47] J. Park, D. Kotzias, R.L. Logan, P. Kuo, K. Merced, S. Singh, M. Tanana, E. Karra-Taniskidou, J. Elston Lafata, D.C. Atkins, M. Tai-Seale, Z.E. Imel, P. Smyth, Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions, J. Am. Med. Inform. Assoc. 26 (2019) 1493–1504.
[48] P.G. Georgiou, M.P. Black, A.C. Lammert, B.R. Baucom, S.S. Narayanan, "That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features? Affective Computing and Intelligent Interaction, (2011), pp. 87–96, doi:http://dx.doi.org/10.1007/978-3-642-24600-5_12.
[49] M. Tanana, A. Dembe, C.S. Soma, Z. Imel, D. Atkins, V. Srikumar, Is sentiment in movies the same as sentiment in psychotherapy? Comparisons using a new psychotherapy sentiment database, Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology (2016) 33–41.
[50] C. Howes, M. Purver, R. McCabe, Linguistic indicators of severity and progress in online text-based therapy for depression, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (2014) 7–16.
[51] J. McAuley, J. Leskovec, Hidden factors and hidden topics, Proceedings of the 7th ACM Conference on Recommender Systems - RecSys' 13 (2013), doi:http://dx.doi.org/10.1145/2507157.2507163.
[52] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment analysis of twitter data, Proceedings of the Workshop on Language in Social Media (LSM 2011) (2011) 30–38.
[53] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: an attentive RNN for emotion detection in conversations, Proceedings of the AAAI Conference on Artificial Intelligence, 33(2019), pp. 6818–6825, doi:http://dx.doi.org/10.1609/aaai.v33i01.33016818.
[54] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: a graph convolutional neural network for emotion recognition in conversation, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019), doi:http://dx.doi.org/10.18653/v1/d19-1015.
[55] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: interactive conversational memory network for multimodal emotion detection, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018), doi:http://dx.doi.org/10.18653/v1/d18-1280.
[56] B. Schuller, M. Valster, F. Eyben, R. Cowie, M. Pantic, AVEC 2012, Proceedings of the 14th ACM International Conference on Multimodal Interaction - ICMI' 12 (2012), doi:http://dx.doi.org/10.1145/2388676.2388776.
[57] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: a multimodal multi-party dataset for emotion recognition in conversations, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), doi:http://dx.doi.org/10.18653/v1/p19-1050.
[58] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359, doi:http://dx.doi.org/10.1007/s10579-008-9076-6.
[59] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016).
[60] M. Tai-Seale, L.A. Hatfield, C.J. Wilson, C.D. Stults, T.G. McGuire, L.C. Diamond, R.M. Frankel, L. MacLean, A. Stone, J. Elston Lafata, Periodic health examinations and missed opportunities among patients likely needing mental health care, Am. J. Manag. Care 22 (2016) e350–e357.
[61] T. Wunderlich, G. Cooper, G. Divine, S. Flocke, N. Oja-Tebbe, K. Stange, Je. Lafata, Inconsistencies in patient perceptions and observer ratings of shared decision making: the case of colorectal cancer screening, Patient Educ. Couns. 80 (2010) 358–363 PMID 20667678.
[62] J.A. Teresi, M. Ramírez, K. Ocepek-Welikson, M.A. Cook, The development and psychometric analyses of ADEPT: an instrument for assessing the interactions between doctors and their elderly patients, Ann. Behav. Med. 30 (2005) 225–242.
[63] J.L. Elman, Finding structure in time, Cogn. Sci. 14 (1990) 179–211, doi:http://dx.doi.org/10.1207/s15516709cog1402_1.
[64] Y. Goldberg, Recurrent neural networks: modeling sequences, in: neural network methods for natural language processing, Synth. Lect. Hum. Lang. Technol. 10 (2017) 163–175, doi:http://dx.doi.org/10.2200/s00762ed1v01y201703hlt037.
[65] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder–decoder approaches, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (2014), doi:http://dx.doi.org/10.3115/v1/w14-4012.
[66] C.D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, (2008), doi:http://dx.doi.org/10.1017/cbo9780511809071.
[67] A. Finset, L. Del Piccolo, Nonverbal Communication in Clinical Contexts, Communication in Cognitive Behavioral Therapy, (2011), pp. 107–128, doi:http://dx.doi.org/10.1007/978-1-4419-6807-4_5.
[68] M. Kayaalp, Patient privacy in the era of big data, Balkan Med. J. 35 (2018) 8–17.
[69] D.C. Dugdale, R. Epstein, S.Z. Pantilat, Time and the patient-physician relationship, J. Gen. Intern. Med. (14 Suppl. 1) (1999) S34–40.
[70] T.R. Konrad, C.L. Link, R.J. Shackelton, L.D. Marceau, O. von dem Knesebeck, J. Siegrist, S. Arber, A. Adams, J.B. McKinlay, It's about time: physicians' perceptions of time constraints in primary care medical practice in three national healthcare systems, Med. Care 48 (2010) 95–100.
[71] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, Interspeech 2019, (2019), doi:http://dx.doi.org/10.21437/interspeech.2019-2680.