78300_Auto_Edited1.docx

WORD COUNT	5223	TIME SUBMITTED	30-SEP-2022 03:50PM
		PAPER ID	91101887

Name of Journal: *World Journal of Psychiatry*

Manuscript NO: 78300

Manuscript Type: REVIEW

Emotion recognition support system: Where physicians and psychiatrists meet linguists and data engineers

Abstract

An important factor in the course of daily medical diagnosis and treatment is understanding patients' emotional states by the caregiver physicians. However, patients usually avoid speaking out their emotions when expressing their somatic symptoms and complaints to their non-psychiatrist doctor. On the other hand, clinicians usually lack the required expertise (or time) and have a deficit in mining various verbal and non-verbal emotional signals of the patients. As a result, in many cases, there is an emotion recognition barrier between the clinician and the patients making all patients seem the same except for their different somatic symptoms.

In particular, we aim to identify and combine three major disciplines (psychology, linguistics, and data science) approaches for detecting emotions from verbal communication and propose an integrated solution for emotion recognition support. Such a platform may give emotional guides and indices to the clinician based on verbal communication at the consultation time.

Key Words: Physician-Patient Relations; Emotions; Verbal behavior; Linguistics; Psychology; Data Science

Adibi P, Kalani SD, Zahabi SJ, Asadi H, Bakhtiar M, Heidarpour MR, Roohafza H, Shahoon H, Amouzadeh M. Emotion recognition support system: Where physicians and psychiatrists meet linguists and data engineers. *World J Psychiatry* 2022; In press

Core Tip: In the context of doctor-patient interactions, we focus on patient speech emotion recognition as a multifaceted problem viewed from three main perspectives: Psychology/psychiatry, linguistics, and data science. Reviewing the key elements and approaches within each of these perspectives, and surveying the current literature on them, we recognize the lack of a systematic comprehensive collaboration among the three disciplines. Thus, motivated by the necessity of such multidisciplinary collaboration, we propose an integrated platform for patient emotion recognition, as a collaborative framework towards clinical decision support.

INTRODUCTION

In order to establish a therapeutic relationship between physician and patient, it is necessary to have knowledgeable practitioners in various specialties as well as an effective interaction and communication between physician and patient which starts with obtaining the patient's medical history and continues to convey a treatment plan[1,2]. Doctor-patient communication is a complex interpersonal interaction where different types of expertise and techniques are required to understand this relationship completely in verbal and nonverbal forms, especially when trying to extract emotional states and determinants during a medical consultation session[3]. Doctor-patient communication is a complex interpersonal interaction which requires an understanding of each party's emotional state. In this paper, our focus is on physicians' understanding of patients' emotions. When patients attend medical consultation, they generally convey their particular experiences of the perceived symptoms to physicians. They interpret these somatic sensations in terms of many different factors including their unique personal and contextual circumstances. Motivated by the illness experience, they generate their own ideas and concerns (emotions), leading them to seek out consultation[4-6]. Generally, patients expect and value their doctors caring for these personal aspects of their experience[7,8]. During interactions and conversations with patients, physicians should be able to interpret their emotional states, which can help build up trust between patients and them[9,10]. This will ultimately lead to better clinical outcomes. Also, identifying and recording these states will help complete patients' medical records. Many diseases that seem to have physical symptoms are, in fact, largely intertwined with psychological variables, such as functional somatic syndromes (FSS)[11]. Increasingly, physicians have realized that recognizing the psychological state of patients with FSS will be very effective in providing an appropriate treatment. For example, the ability to accurately understand sound states may help interpret a patient's pain. Thus, the presence of information about patients' mental states in their medical records is essential.

Emotion detection accuracy, *i.e.*, the ability to detect whether a patient is expressing an emotion cue, has consequences for the physician–patient relationship. The key to patient-centered care is the ability to detect, accurately identify, and respond appropriately to the patient's emotions[12-15]. Failure to detect a patient's emotional cues may give rise to an ineffective interaction between doctor and patient, which may, in turn, lead to misdiagnosis, lower recall, mistreatments, and poorer health outcomes[16,17]. Indeed, if the emotion cue is never detected, then the ability to accurately identify or respond to the emotion never comes into play. Doctors who are more aware of their patients' emotions are more successful in treating them[13]. Patients have also reported greater satisfaction with such physicians[18-22]. Recognizing the emotions and feelings of patients provides the ground for more physician empathy with patients[23,24]. The academic and medical literature highlights the positive effects of empathy on patient care[25]. In this regard, the medical profession requires doctors to be both clinically competent and empathetic toward the patients. However, in practice, meeting both needs may be difficult for physicians (especially inexperienced and unskilled ones)[26]. On the other hand, patients do not always overtly express these experiences, feelings, concerns, and ideas. Rather, they often communicate them indirectly through more or less subtle nonverbal or verbal "clues" which nevertheless contain interesting clinical information which can be defined as "clinical or contextual clues"[27-29]. They do not say, ''Hey doctor, I'm feeling really emotional right now; or do you know whether I'm angry or sad?'' Thus, emotional cues are often ambiguous and subtle[30-33].

On the other hand, patients' emotional audiences (*i.e.*, physicians) are often inexperienced in detecting emotions. One of the most important problems physicians face in the development of this process is the difficulty of capturing the clues that patients offer and failing to encourage them to expose details about these feelings[34]. Research indicates that over 70% of patients' emotional cues are missed by physicians[34]. It is unclear whether missed responses were the result from physicians detecting an emotional cue and choosing not to respond, or from failing to detect the cue in the first place. Indeed, these emotional cues present a challenge to doctors who often overlook them, as clinical information and therefore opportunities to know the patient's world are lost[34-37]. Physicians vary in their ability to recognize patients' emotions, with some being fully aware of the significance of understanding emotions and capable of identifying them. They also range from high emotional intelligence to low emotional intelligence. Another argument often heard from physicians is that they do not have time for empathy[38].

Despite the importance of such issues, this aspect remains grossly overlooked in conventional medical training. This comes from the fact that training emotion skills in medical schools is variable, lacks a strong evidence- base, and often does not include the training of emotion processing[39].

In the preceding paragraphs, four reasons were offered as to why physicians have failed to detect and interpret patients' emotional states, and hence why we need to find a solution for this problem. These reasons could be summarized as follows. First, detecting patients' emotions can contribute to healing them, as well as to increasing their satisfaction. Secondly, emotional cues are mostly indirectly found in patients' speech. That is, emotional cues can be very subtle and ambiguous. Further, many physicians do not possess enough experience to detect patients' emotions or even when they are skilled and experienced enough to do so, they do not have time to deal with it. In addition, training doctors to detect patients' emotions has been thoroughly overlooked in routine medical training. Thus, if a solution can be found to help physicians recognize patients' emotions and psychological states, this problem can be overcome to a large extent.

One strategy is to develop and employ a technology that can provide information about the patient's emotions, feelings, and mental states by processing their verbal and non-verbal indicators (See Figure 1). In the present manuscript, we focus on verbal communication. Human speech carries a tremendous number of informative features, which enables listeners to extract a wealth of information about speakers' identity. These features can range from linguistic characteristics through extralinguistic features to paralinguistic information, such as the speaker's feelings, attitudes, or psychological states[40]. The psychological states (including emotions, feelings, and affections) embedded in people's speech are among the most important parts of the verbal communication array humans possess. As with other non-verbal cues, they are under conscious control much less than verbal cues. This makes speech an excellent guide to a human's "true" emotional state even when he/she is trying to hide it.

In order to design and present such technology, the first step is to know which indicators in speech can be used to identify emotions. Psychologists, psychiatrists, and linguists have done extensive research to identify people's emotions and feelings, and have identified a number of indicators. They believe that through these markers, people's emotions and feelings can be understood.

THE PSYCHOLOGICAL APPROACH

Psychologists and psychiatrists pay attention to content indicators and acoustic variables to identify people's emotions through their speech. Scholarly evidence suggests that mental health is associated with specific word use[41-43]. Psychologists and psychiatrists usually consider three types of word usage to identify emotions: (1) positive and negative emotion words; (2) standard function word categories; and (3) content categories. They distinguish between positive ("happy", "laugh") and negative ("sad", "angry") emotion words, standard.

Figure 1 Emotion indicators in the patient-doctor interaction

function word categories (*e.g.*, self-references, first, second, and third person pronouns) and various content categories (*e.g.*, religion, death, and occupation). The frequent use of "You" and "I" suggests a different relationship between the speaker and the addressee than that of "We". The former suggests a more detached approach, whereas the latter expresses a feeling of solidarity. Multiple studies have indicated that the frequent use of the first-person singular is associated with negative affective states[44-48], which reveals a high degree of self-preoccupation[49]. People with negative emotional states (such as sadness or depression) use second and third person pronouns less often[38-40]. These people have a lower ability to express positive emotions and express more negative emotions in their speech[44-48]. Also, people with negative emotional states use more words referring to death[44].

In addition to the content of speech, psychologists and psychiatrists also look at several acoustic variables (such as pitch variety, pause time, speaking rate, and emphasis) to detect emotions. According to the research in this area, people with negative emotional states typically have a slower speaking rate [50-54], lower pitch variety[55,56], produce fewer words[57], and have longer pauses[53,54,58].

THE LINGUISTIC APPROACH

Within linguistics, various approaches (*e.g.*, phonetic, semantic, discourse-pragmatic, and cognitive) have been adopted to examine the relationship between language and emotion[56,60]. As far as the phonetic and acoustic studies are concerned, emotions can be expressed through speech and are typically accompanied with physiological signals such as muscle activity, blood circulation, heart rate, skin conductivity, and respiration. This will subsequently affect the kinematic properties of the articulators, which in turn will cause altered acoustic characteristics of the produced speech signals of the speakers. Studies of the effects of emotion on the acoustic characteristics of speech have revealed that parameters related to the frequency domain (*e.g.*, average values and ranges of fundamental frequency and formant frequencies), the intensity domain of speech (*e.g.*, energy, amplitude), temporal characteristics of speech (*e.g.*, duration and syllable rate), spectral features Mel frequency cepstral coefficients, and voice quality features (*e.g.*, jitter, shimmer, and harmonics-to-noise-ratio are amongst the most important acoustically measurable parameters for correlates of emotion in speech. For instance, previous studies have reported that the mean and range of fundamental frequency observed for utterances spoken in anger situations were considerably greater than the mean and range for the neutral ones, while the average fundamental frequency for fear was lower than that observed for anger[61] (see Figure 2 & Table 1).

Past research has produced many important findings to indicate that emotions can be distinguished by acoustical patterns; however, there are still a multitude of challenges regarding emotional speech research. One of the major obstacles that must be tackled in the domain of emotion recognition relates to variable vocalization which exists within speakers. Voices are often more variable within the same speaker (within-speaker variability) than they are between different speakers and it is thus unclear how human listeners can recognize individual speakers' emotion from their speech despite the tremendous variability that individual voices reveal. Emotion is sensitive to a large degree of variation within a single speaker and is highly affected by factors such as gender, speakers, speaking styles, sentence structure in spoken language, culture, and environment. Thus, identifying what specific mechanisms motivate variability in acoustic properties of emotional speech and how we can overcome differences arising from individual properties remain major challenges ahead of the emotion recognition field.

With regard to investigations in the area of pragmatics (in its continental notion which encompasses discourse analysis, sociolinguistics, cognitive linguistics, and even semantics), we observe a flourishing trend in linguistics focusing on the emotion in language[59,62]. These studies have examined important issues related to referential and non-referential meanings of emotion. In semantics, the focus has been on defining emotional and sentimental words and expressions, collocations and frames of emotion[63,64], field semantics[62], as well as lexical relations including semantic extensions. However, more pragmatic and discourse-oriented studies have looked at issues in terms of emotion and cultural identity[65,66]; information structure/packaging (*e.g.* topicalization and thematicization[67] and emotion, emotive particles and interjections[68-70], emotional implicatures, and emotional illocutionary acts, deixis, and indexicality (*e.g.* proximalization and distalization[71,72], conversational analysis and emotion (*e.g.* turn-taking and interruption)[73,74], *etc.*

Figure 2 Spectrograms of the Persian word [xanom] pronounced by a Persian male speaker in neutral (top) and anger (down) situations. Figure 2 shows spectrograms of the word [xanom], spoken by a native male speaker of Persian. The figure illustrates a couple of important differences between acoustic representations of the produced speech sounds. For example, the mean fundamental frequency in anger situations is higher (161 Hz) than that observed for neutral situations (112 Hz). Additionally, acoustic features such as mean formant frequencies (*e.g.* F1, F2, F3, and F4), minimum and maximum of the fundamental frequency, and mean intensity are lower in neutral situations. More details are provided in Table 1.

Cognitive linguists use other methods to recognize emotion in speech. The cognitive linguistic approach to emotion concepts is based on the assumption that conventionalized language used to talk about emotions is a significant tool in discovering the structure and content of emotion concepts[75]. They consider a degree of universality for emotional experience and hold that this partial universality arises from basic image schemas that emerge from fundamental bodily experiences[76-79]. In this regard, the cultural model of emotions is a joint product of (possibly universal) actual human physiology, metonymic conceptualization of actual human physiology, metaphor, and cultural context[77]. In this approach, metaphor and metonymy are used as conceptual tools to describe the content and structure of emotion concepts.

Table 1 Acoustic differences related to prosody and spectral features of the word [xanom] produced by a Persian male speaker in neutral and anger situations.

Conceptual metaphors create correspondences between two distinct domains. One of the domains is typically more physical or concrete than the other (which is thus more abstract)[76]. For example, in the Persian expression *gham dar delam âshiyâneh kardeh* 'sadness has nested in my heart', *gham* 'sadness' is metaphorically conceptualized as a bird and *del* 'heart/stomach' is conceived of as a nest. The metaphor focuses on the perpetuation of sadness. The benefit of metaphors in the study of emotions is that they can highlight and address various aspects of emotion concepts[75,76]. Metonymy involves a single domain, or concept. Its purpose is to provide mental access to a domain through a part of the same domain (or vice versa) or to a part of a domain through another part in the same domain[80]. Metonymies can express physiological and behavioral aspects of emotions[75]. For example, in *she was scarlet with rage*, the physiological response associated with anger, *i.e.*, redness in face and neck area, metonymically stands for anger. Thus, cognitive linguistics can contribute to the identification of metaphorical and metonymical conceptualizations of emotions in large corpora.

Although speech provides substantial information about the emotional states of speakers, accurate detection of emotions may nevertheless not always be feasible due to challenges that pervade communicative events involving emotions. Variations at semantic, pragmatic, and social-cultural levels present challenges that may hinder accurately identifying emotions *via* linguistic cues. At the semantic level, one limitation seems to be imposed by the "indeterminacy of meaning", a universal property of meaning construction which refers to "situations in which a linguistic unit is underspecified due to its vagueness in meaning"[81]. For example, Persian expressions such as *ye juriam* or *ye hâliam* roughly meaning 'I feel strange or unknown' even in context may not explicitly denote the emotion(s) the speaker intends to convey, and hence underspecify the conceptualizations that are linguistically coded. The other limitation at the semantic level pertains to cross-individual variations in the linguistic categorization of emotions. Individuals differ as to how they linguistically label their emotional experiences. For example, the expression *tu delam qoqâst* 'there is turmoil in my heart' might refer to 'extreme sadness' for one person but might suggest an 'extreme sense of confusion' for another. Individuals also reveal varying degrees of competence in expressing emotions. This latter challenge concerns the use of emotion words, where social categories such as age, gender, ethnic background, education, social class, and profession could influence the ease and skill with which speakers speak of their emotions. Since emotions perform different social functions in different social groups[82], their use is expected to vary across social groups.

Language differences are yet another source of variation in the use and expression of emotions, which presents further challenges to the linguistic identification of emotions. Each language has its own specific words, syntactic structures, and modes of expressions to encode emotions. Further, emotions are linked with cultural models and reflect cultural norms as well as values[83]. Thus, emotion words cannot be taken as culture-free analytical tools or as universal categories for describing emotions[84]. Patterns of communication vary across and within cultures. The link between communication and culture is provided by a set of shared interpretations which reflect beliefs, norms, values, and social practices of a relatively large group of people[85]. Cultural diversity may pose challenges to doctors and health care practitioners in the course of communicating with patients and detecting their emotions. In a health care setting, self-disclosure is seen as an important (culturally sensitive) characteristic that differentiates patients according to their degree of willingness to tell the doctor/practitioner what they feel, believe, or think[86]. Given the significance of self-disclosure and explicitness in the verbal expression of feelings in health care settings (Robinson, ibid), it could be predicted that patients coming from social groups with more indirect, more implicit, and emotionally self-restrained styles of communication will probably pose challenges to doctors in getting them to speak about their feelings in a detailed and accurate manner. In some ethnic groups, self-disclosure and intimate revelations of personal and social problems to strangers (people outside one's family or social group) may be unacceptable or taboo due to face considerations. Thus, patients belonging to these ethnic groups may adopt avoidance strategies in their communication with the doctor and hide or understate intense feelings. People may also refrain from talking about certain diseases or use circumlocutions due to the taboo or negative overtones associated with them. Further, self-restraint may be regarded as a moral virtue in some social groups, which could set a further obstacle in self-disclosing to the doctor or healthcare practitioner.

Overall, it is seen that these linguistically-oriented studies reveal important aspects of emotion in language use. In particular, they have shown how emotion is expressed and constructed by speakers in discourse. Such studies, however, are not based on multi-modal research to represent a comprehensive and unified description of emotion in language use. This means that, for a more rigorous and fine-grained investigation, we need an integrative and cross-disciplinary approach to examining emotions in language use.

THE DATA SCIENCE APPROACH

From the data science perspective, speech emotion recognition (SER) is a machine learning (ML) problem whose goal is to classify the speech utterances based on their underlying emotions. This can be viewed from two perspectives: (1) Utterances as sounds with acoustic and spectral features (non-verbal); and (2) Utterances as words with specific semantic properties (verbal)[87-91]. While in the literature, SER typically refers to the former perspective, the latter is also important and provides a rich source of information, which can be harvested in favor of emotion recognition *via* natural language processing (NLP). Recent advances in the NLP technology allow for a fast analysis of text. In particular, word vector representations (also known as word embeddings) are used to embed words in a high dimensional space where words maintain semantic relationships with each other[92]. These vector representations, which are obtained through different ML algorithms, commonly capture the semantic relations between the words by looking into their collocation/co-occurrence in large corpora. In this way, the representation of each word and the machine's understanding of that partially reflect the essential knowledge that relates to that word, thus capturing the so-called frame semantics. The problem of SER can thus be tackled by analyzing the transcript of the speech by running various downstream tasks on the word vectors of the given speech.

As for the former perspective, different classifiers have so far been suggested for SER as candidates for a practically feasible automatic emotion recognition (AER) system. These classifiers can be put broadly into two main categories: Linear classifiers and non-linear classifiers. The main classification techniques/models within these two categories are:

Hidden Markov model[93-96]

Gaussian mixture model[97,98]

K-Nearest neighbor[99]

Support vector machine[100,101]

Artificial neural network[94,102]

Bayes classifier[94]

Linear discriminant analysis[103,104]

Deep neural network[102-107]

A review of the most relevant works within the above techniques has recently been done in[108]. We have provided a short description of the above techniques in Appendix. One of the main approaches in the last category, *i.e.*, deep neural networks, is to employ transfer learning. Recently[109] has reviewed the application of generalizable transfer learning in AER in the existing literature. In particular, it provides an overview of the previously proposed transfer learning methods for speech-based emotion recognition by listing 21 relevant studies.

The classifiers developed for SER may also be categorized in terms of their feature sets. Specifically, there are three main categories of speech features for SER:

(a) the prosodic features[110-112]

(b) the excitation source features[110,111,115,116]

(c) the spectral or vocal tract features[117-120]

Table 2 Different approaches to recognizing the emotional indicators in speech.

Prosodic features, also known as continuous features, are some attributes of the speech sound such as pitch or fundamental frequency and energy. These features can be grouped into the following subcategories[104,105]: (1) pitch-related features; (2) formant features; (3) energy-related features; (4) timing features; and (5) articulation features. Excitation source features, which are also referred to as voice quality features, are features which are used to represent glottal activity, such as harshness, breathiness, and tenseness of the speech signal.

Finally, spectral features, also known as segmental or system features, are the characteristics of various sound components generated from different cavities of the vocal tract system that have been extracted in different forms. The particular examples are ordinary linear predictor coefficients[117], one-sided autocorrelation linear predictor coefficients[113], short-time coherence method[114], and least squares modified Yule–Walker equations[115].

Given the breadth and complexity of emotion detection indicators in psychology and linguistics, it is difficult to establish a decision support system for a doctor's emotional perception of patients. This requires a comprehensive and multidisciplinary approach. In order to build such a system, an application will be very useful. When a person experiences intense excitement, in addition to a reduction in his/her concentration, his/her mental balance is also disturbed more easily and quickly. This is also used as a strategy in sociology to take hold of people's minds.

Under unstable conditions, reasoning and logical thinking (and thus more effective and active behavior), which emerge in response to the activity of new and higher parts of the brain, are dominated by older parts of the brain, which have more biological precedents (several thousand *vs* millions of years). Thus, these older parts act impulsively or reactively.

Working in an emergency environment and sometimes even in an office has special conditions, such as excessive stress due to medical emergencies, pressure from patient companions, patient's own severe fear, as well as the impact of the phenomenon of "transference" and "countertransference" between physician and patient or between physician and patient companion. These can impair a physician's ability to reason and think logically. Thus, use of such an intelligent system can enhance doctors' efficiency, increase their awareness, and make it easier for them to manage the conditions.

THE PROPOSED SOLUTION

In the previous sections, the problem of SER was viewed from its three main perspectives: Psychology/psychiatry, linguistics, and data science, and the key elements within each perspective were highlighted. One way to integrate these three sides and benefit from their potential contributions to SER is through developing an intelligent platform. In what follows, focusing on SER in the context of doctor-patient interactions, we propose a solution for such integration.

The proposed solution consists of two key components:

The intelligent processing engine

The data-gathering platform

The intelligent processing engine, at the algorithmic level, is based on NLP, speech processing, and in a wider context, behavioral signal processing methods. While it is clear that the processing engine will serve as the brain of the proposed intelligent platform, and is indeed a place where the novelty, creativity, and robustness of implemented algorithms can make a great difference, it will not practically function desirably without a well-thought, flexible data-gathering platform. Thus, despite the genuine algorithms which are to be developed at the core of the platform, and the undeniable impact they will have on the performance of the system, we believe it is the data-gathering platform that will make the solution very unique. One idea is to develop a cloud-based multi-mode multi-sided data gathering platform, which has three main sides:

The patient side

The physician side

The linguistic/psychologist side

Regarding the functioning of the platform, three modes can be considered:

The pre-visit mode

The on-visit mode

The post-visit mode

The pre-visit mode will include the patient's declaration of his/her health-related complaints/conditions and concerns, which will be automatically directed to the cloud-based processing engine, and labeled *via* a speech-emotion recognition (SER) algorithm. This mode is reinforced *via* receiving additional multi-dimensional data from the patient through filling various forms and questionnaires. Also, it is possible for the patient to submit text to accompany his/her speech. This allows one to perform additional classification/clustering tasks such as sentiment analysis or patient segmentation on the provided text, using biomedical NLP methods. The on-visit mode enables the recording of the visiting session and the clinician-patient conversations. Finally, the post-visit mode of the application provides an interface for the psychiatrist/psychologist as well as the linguist to extract and label the psychological and linguistic features within the patient’s speech. Such tagging of the data by a team of specialists will in the long term lead to a rich repository of patient speech, which is of great value in training the ML algorithms in the processing engine.

Figure 3 Integrated platform for patient emotion recognition and decision support (INDICES). It consists of the data gathering platform and the intelligent processing

engines. Each patient's data, in the form of voice/transcripts is captured, labeled, and stored in the dataset. The resulting dataset feeds the machine language training/validation and test engines. The entire process of intelligent processing may iterate several times for further fine tuning. It is crucial to have collaboration among the three relevant expertise in different parts of the proposed solution.

Although the proposed platform is to be designed such that it scales up at the population level in order to benefit from the diversity of the gathered data, it will also serve every individual as a customized personalized electronic health record that keeps track of the patient's psycho-emotional profile. As for the implementation of the platform, it is practically possible to tailor it to various devices (cell phones, tablets, PCs, and Laptops) *via* android/macOS, and web service applications

Note that emotion is essentially a multifaceted concept and no matter how sophisticated the proposed signal processing and data mining technology is, it would eventually face limitations in grasping all of its aspects. For instance, cultural aspects of expressing emotions can be a serious challenge to the technological system. Extracting the appropriate measurable features for correctly interpreting the cultural indices of emotion in speech can be a challenge, which nonetheless adds to the beauty of the problem. Further, as mentioned earlier, not all emotional indicators are embedded in the speech. Indeed, facial expressions and body gestures play important roles in expressing one's emotions as well. Hence, since the technology considered in our proposed method is focused merely on speech signals, it will of course have *blind spots* such as the visual aspects of emotion which are not exploited. This can be thought of as a main limitation that bounds the performance of the proposed emotion recognition system. However, with the same pattern that technology has always emerged throughout history, the proposed method can similarly serve as a baseline to which further improvements and additional capabilities can be added in future. We must also note that in capturing the different aspects of emotion, we are faced with a tradeoff between computational complexity and performance. In particular, depending on the required accuracy of the system, one may need to customize the aspects of emotion which are to be examined *via* technology, taking into account the computational burden they would impose on the system.

We shall finally end this section with two remarks. First, it is important to note that despite all integrations and optimizations involved in the design and training of the proposed intelligent platform, it would still have the intrinsic limitations of a machine as a decision-maker, some of which were mentioned above. Thus, the proposed solution would eventually serve as a decision aid/support (and not as a decision replacement). Secondly, while the proposed solution provides a global framework, it invites for a series of methodologies and solutions, which are to be adapted and customized to each language and culture setting for local use.

APPENDIX

Here we provide the following table which includes a brief description of each of the data science techniques and models mentioned earlier, along with reference sources in which further technical details of the methods can be found.

Table 3 A brief description of some data science models/methods.

CONCLUSION

In the context of doctor-patient interactions, this article focused on patient SER as a multidimensional problem viewed from three main aspects: Psychology/psychiatry, linguistics, and data science. We reviewed the key elements and approaches within each of these three perspectives, and surveyed the relevant literature on them. In particular, from the psychological/psychiatric perspective, the emotion indicators in the patient-doctor interaction were highlighted and discussed. In the linguistic approach, the relationship between language and emotion was discussed from phonetic, semantic, discourse-pragmatic, and cognitive perspectives. Finally, in the data science approach, SER was discussed as a ML/signal processing problem. The lack of a systematic comprehensive collaboration among the three discussed disciplines was pointed out.

Motivated by the necessity of such multidisciplinary collaboration, we proposed a platform named INDICES: An integrated platform for patient emotion recognition and decision support. The proposed solution can serve as a collaborative framework towards clinical decision support.

78300_Auto_Edited1.docx

ORIGINALITY REPORT

10%
SIMILARITY INDEX

PRIMARY SOURCES

1	creativecommons.org Internet	120 words — 2%
2	Blanch-Hartigan, Danielle. "Patient satisfaction with physician errors in detecting and identifying patient emotion cues", Patient Education and Counseling, 2013. Crossref	119 words — 2%
3	dokumen.pub Internet	61 words — 1%
4	ceur-ws.org Internet	56 words — 1%
5	pure.uvt.nl Internet	56 words — 1%
6	www.ncbi.nlm.nih.gov Internet	28 words — 1%
7	link.springer.com Internet	22 words — < 1%
8	bmcmedethics.biomedcentral.com Internet	21 words — < 1%
9	Alapan Bandyopadhyay, Sarbari Sarkar, Abhijit Mukherjee, Sharmistha Bhattacherjee, Soumya	18 words — < 1%

	Basu. "Identifying emotional Facial Expressions in Practice: A Study on Medical Students", Indian Journal of Psychological Medicine, 2020 Crossref	
10	ujcontent.uj.ac.za Internet	17 words — < 1%
11	worldwidescience.org Internet	15 words — < 1%
12	Martin J.H. Balsters, Emiel J. Krahmer, Marc G.J. Swerts, Ad J.J.M. Vingerhoets. "Verbal and Nonverbal Correlates for Depression: A Review", Current Psychiatry Reviews, 2012 Crossref	12 words — < 1%
13	www.frontiersin.org Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS