Designing Computational Tools for Behavioral and Clinical Science

Albert Ali Salah
Utrecht University
Department of Computing and Information Sciences
Utrecht, the Netherlands
Boğaziçi University
Department of Computer Engineering
Istanbul, Turkey
a.a.salah@uu.nl

ABSTRACT

Automatic analysis of human affective and social signals brought computer science closer to social sciences and, in particular, enabled collaborations between computer scientists and behavioral scientists. In this talk, I highlight the main research areas in this burgeoning interdisciplinary area, and provide an overview of the opportunities and challenges. Drawing on examples from our recent research, such as automatic analysis of interactive play therapy sessions with children, and diagnosis of bipolar disorder from multimodal cues, as well as relying on examples from the growing literature, I explore the potential of human-AI collaboration, where AI systems do not replace, but support monitoring and human decision making in behavioral and clinical sciences.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; **Psychology**; • **Computing methodologies** → *Machine learning*; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Affective computing, social computing, human behavior understanding

ACM Reference Format:
Albert Ali Salah. 2021. Designing Computational Tools for Behavioral and Clinical Science. In *Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '21 Companion), June 8–11, 2021, Virtual Event, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3459926.3464906

1 INTRODUCTION

AI based systems offer new possibilities in behavioural and clinical sciences. New and improved sensors, combined with advanced digital signal processing technologies make new measurements possible. Pattern recognition and machine learning approaches allow the construction of classifiers of ever increasing complexity, where multimodal sensory signals are connected to informative indicators. Some of the qualitative observation methods with well-known shortcomings can be supplemented with quantitative approaches. For example, subtle indicators can be observed over long periods to accumulate evidence, and unreliable and repeated self-report measures can be with validated with quantitative data.

Obviously, there are also challenges that need to be overcome. Computational approaches are often data-driven, and statistical in nature. Apart from requiring large amounts of data for proper training, they also have serious problems dealing with outliers and rare cases. It is not so difficult to design a computational experiment poorly, so that the results are overly optimistic. Peeking at the test set accuracy during model selection is all it takes. The richness of the model space is both a blessing and a curse; with so many modeling choices, and ever more complex models, out of sample generalization becomes an issue.

In almost all interdisciplinary endeavors, experienced researchers warn us that language gap between disciplines needs to be bridged first. In clinical sciences, choosing a representative sample is very important, a lot of factors and demographics are observed and controlled. In computational sciences, controlling model complexity and generalization are very important. These concerns do not clash, but they are prioritized to a different extent in these disciplines. Furthermore, computational sciences are more result-driven, we often see accuracy and precision of the models being more important than, say, the explainability and transparency. Even with newer tools of visualization, deep neural network based modeling approaches remain opaque to their end-user; the network visualizations are primarily for providing insights to the developers. Subsequently, there is a need for explainable models that facilitate clinician's use of such computational tools. If one considers a scale of observations, from low-level and clearly defined quantities (e.g. the temperature of a child) to more elaborate and hard to quantify indicators (e.g. the stress level of a child), it would be safe to say the need for explainability rises with the complexity of the indicator.

This brief contribution presents an overview of concerns and directions for this area, drawing heavily on my past research for illustrative examples. I define a taxonomy in the next section that bridges affective and social computing, and discuss application areas of computational tools for behavioral and clinical sciences. Section 3 and Section 4 are two case studies, on affect analysis in

EICS '21 Companion, June 8–11, 2021, Virtual Event, Netherlands.

ACM ISBN 978-1-4503-8449-0/21/06…$15.00
https://doi.org/10.1145/3459926.3464906

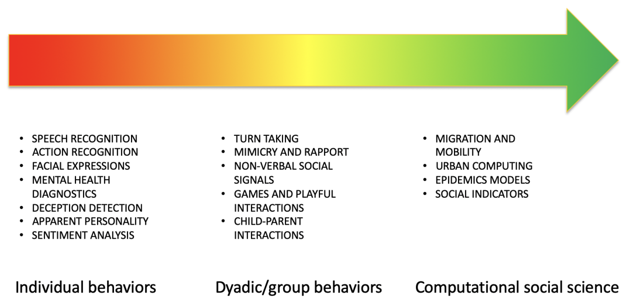

Figure 1: Applications of social and affective computing, depending on the number of people involved in the analysis.

children's play therapy and multimodal analysis of bipolar disorder, respectively. Section 5 concludes the paper.

2 APPROACHES IN AFFECTIVE AND SOCIAL COMPUTING

Computer analysis of human behavior starts with sensing behaviors. This can be performed with physical sensors, such as cameras, microphones, wearable sensors, or sensors on mobile phones. It is also possible to sense the virtual behaviors, for example behaviors on social media, on Internet based games, and mobile phone usage. Both physical and virtual sensors deliver large amounts of data over human behavior, and over longer periods.

Once the sensing is accomplished, the data will be analyzed. Theory-driven analysis approaches of human behavior can be found in cognitive science, psychology, and sociology, depending on the resolution of the analysis. These approaches start from an hypothesis, and examine data to support or reject the hypothesis. Data-driven approaches are typically based on pattern recognition, machine learning, and to some extent, statistics. In many cases, the analysis is performed in a black-box fashion, and supervised classification tasks or unsupervised clustering tasks are defined. The main drawback of the latter approach is that spurious patterns can be detected easily in this manner, and over-analysis is a common issue.

There are many domains for human behavior analysis, such as multimedia analysis, surveillance, ambient intelligence, urban computing, and healthcare [28]. An important taxonomy is the temporal resolution of the observed behavior, which can be as fast as an eye-blink in a driver fatigue estimation application, or as slow as the sleep cycles observed over months [29]. Another taxonomy involves the number of observed people. Figure 1 provides examples that span the spectrum from this perspective. Looking at a single individual, possible tasks are speech recognition, action recognition, facial expression analysis, etc. When two individuals (i.e. a dyad) or a small group of individuals are involved, social signals such as turn taking, mimicry, and interactions can be analysed [22, 25, 34]. When behavior of hundreds or thousands of people are analyzed at the same time, we talk of social physics, social computing, or computational social science [18].

In a sense, what brings together these wide range of applications is the complexity of human behavior, which nonetheless harbors some order that can be leveraged to classify or predict it [27], or to design applications and systems to steer and change it [20]. For behavioral and clinical sciences, the application of such technology is very direct, but surprisingly varied. They don't only focus on the analysis of the individual behavior or an interaction from multiple modalities [14, 33], but also include analysis of the context of behavior, which can be culturally-conditioned, or highly domain-specific [27]. For example, designing applications for monitoring elderly subjects brings specific usability challenges, and the cultural context plays a very important role in how a specific technology is perceived [10, 23, 30].

We have summarized the main application areas of social and affective computing for behavioral and clinical sciences in [27] as follows:

- Automated coding: Finding a suitable abstraction of behaviors, and providing transcripts or descriptions.
- Indexing, search, and retrieval: This is the primary usage of analysis technology for multimedia applications, but also useful for archival analyses.
- Quality assessment: Data acquisition quality, as well as data biases can be assessed automatically.
- Diagnosis and prediction: Automatic estimation of indicators, and classification of behaviors.
- Longitudinal analysis: Manual analysis of longitudinal data is costly, tedious, and error-prone. This is one of the most promising areas of automatic analysis.
- Training and simulation: Training the human experts benefits from automatic tools, including newer technologies such as augmented and virtual reality.

As can be seen from the list, analysis tools are varied and can be serve the domain experts at different points. The risks involved in the usage of such technologies are similarly varied. Proper performance assessment for clinical applications is the most basic concern, which is exacerbated by ethnic and demographic biases that come from training sets with limited variance. In clinical scenarios, rare cases and outliers, as well as severe class imbalance is an issue for classifiers. For applications that directly interact with patients, interface issues and handling trust become important points. For systems that are supposed to help diagnosis, computer based approaches are not robust enough for almost all but the most basic tasks. Medical experts rely on cues at different semantic levels and can interpret unexpected, rarely seen, distantly related indicators. Nonetheless, over-reliance on technology has been shown to become an issue in other domains where AI is incorporated into decision making processes, and hard lessons are learned. Explainable AI and more advanced reasoning systems are now being pursued to provide more insight to the decisions reached by AI systems [9].

In the next two sections, I will give specific examples from two different case studies to illustrate some of the challenges and the potential of computational tools.

3 CASE STUDY: AFFECT ANALYSIS IN PLAY THERAPY

The first case study is a collaboration with Prof. Sibel Halfon from the Istanbul Bilgi University Psychological Center, where we have sought to build analysis tools for longitudinally assessing affective states of children, playing games in a room with a psychotherapist [8, 12, 13]. These children were 4-10 years old, and were referred to the clinic due to internalizing and externalizing emotion regulation problems such as rule-breaking, aggression, anxiety, and social problems [11]. They were recorded with two cameras during psychodynamic play therapy with an expert, and weekly sessions of about 50 minutes continued for 40 sessions over a ten-month period, on average, for each child.

The clinical assessment approach uses clinically validated tools, such as the Child Behavior Checklist (CBCL) [2] to identify problematic behaviors, and Children's Play Therapy Instrument (CPTI) [17], which rates children's play activity from multiple dimensions. The speech of the child and the therapist during the sessions are meticulously transcribed, affective and behavioral cues are manually coded by watching hundreds of hours of video. Multiple expert annotators work on the material, and annotator agreement is assessed.

Automatic analysis of this kind of data can have several aims. One of the questions we ask is whether our systems can predict some of the indicators derived from the clinical tools automatically. It is highly desirable to process the data to visualize indicators, such as overall affect over the sessions, as an overview to the therapist. Furthermore, near real-time feedback to the therapist about the interaction quality would be very useful. Automatic speech recognition can be used for replacing the manual transcriptions with computer-transcribed text, saving hundreds of hours. Affective behaviors (e.g. smiles, bursts of anger, etc.) can be automatically detected to enrich the transcripts.

Each of the data channels we deal in this problem faces some non-ideal conditions. Affect analysis can be performed multimodally, using facial expressions [31], text transcriptions [3], as well as paralinguistic cues from the child's speech [16]. The face images are difficult to obtain from the two static cameras, they are often occluded, and there is motion blur [8]. Affect analysis from text faces a completely different challenge; text analysis tools are not very advanced for the Turkish language, which is the native language of the children. Finally, paralinguistic analysis is difficult, because the recording conditions are not ideal, and there is significant noise. While the number of subjects is small for training large computational models, which is typical for many clinical applications, transfer learning is effectively used [14].

The analysis of the results illustrates that different modalities complement each other for different aspects of the task. To predict the pleasure dimension, face and text based affect analysis is used. For other affect classes (i.e. anger, anxiety and sadness), text based affect analysis outperforms face based analysis. This is not surprising, because facial cues are more subtle, and there is far more data loss for this modality.

While the current state of the analysis provides good insights and additional verification for the clinician, automatic analysis can also be employed to quantify the effects of the interventions of the therapist on the children's affective states. Additionally, synchrony, mimicry and rapport can be assessed. Games and play provide excellent opportunities to capture affective displays in ecologically valid conditions [7], and further research can also help improve the technology in these areas.

4 CASE STUDY: MULTIMODAL ANALYSIS OF BIPOLAR DISORDER

Mood disorders are linked to affective states, and have high prevalence. Bipolar disorder (BD) is a major challenge, with low remission rates and treatment compliance. My second case study is a collaboration with doctors Elvan Çiftçi and Hüseyin Güleç, who collected a multimodal bipolar disorder database to investigate whether automatic analysis tools can be used to predict mania levels and remission in BD. The Turkish audio-visual BD corpus is the first of its kind [4], and its aim is to find biological markers of treatment response that can be automatically detected to reduce treatment resistance. In 2018, we have opened this database to the larger research community in form of a multimedia challenge [26], and many teams have had the opportunity to work on the data since then.

The corpus was collected from 51 patients who had manic episodes, and were admitted to the hospital. According to the protocol we have devised, videos of the patients were recorded during several tasks, including explaining the reason the patient came to the hospital, describing a happy and a sad memory, a neutral counting task and its faster version, and describing two emotion-eliciting pictures, respectively. These tasks were performed on the 0th- 3rd- 7th- 14th- 28th day, and after discharge on the 3rd month. In parallel, depressive and manic features were evaluated using Young Mania Rating Scale (YMRS) [36] and Montgomery-Åsberg Depression Rating Scale (MADRS) [21]. Just like the previous case study, ethical committee approvals were obtained and patients gave informed consent for recording and subsequent analyses.

Computer analysis of bipolar patients was not common, because of the difficulty of accessing audio-visual data of such patients. The first challenge was obtaining data and consent for research. Surprisingly, the patients were extremely cooperative. They were suffering greatly under their condition, and they wanted to help with research. Furthermore, there was already prior work on depression when we started this project, and techniques developed in such a similar pathology were highly relevant in the BD context as well [5, 6, 15, 24].

In terms of analysis, similar to the previous case study, we have seen that multimodality was beneficial in terms of estimation accuracy. An important caveat in designing classification (or regression) experiments with a specific pathology is that the negative class cannot be composed of healthy subjects. This is an error frequently made in the computing literature. The negative class should be composed of a variety of pathologies that can act as confounding conditions in the analysis for a realistic experimental setting. However, it is very difficult to collect such a dataset, especially because the tasks are specific to the pathology. Subsequently, in the AVEC challenge [26], we opted for the analysis of mania levels, and did not include healthy controls in the dataset. The accuracies obtained with complex, multi-level and multimodal systems for this task did not reach levels that could warrant clinical applications. What is more important, the better models were using deep neural network estimators, and had millions of parameters. It was difficult to let the clinician know what kind of cues were leveraged to produce a particular estimation.

5 CONCLUSIONS

Classification and analysis of mental disorders is difficult, but computer analysis of multimodal features is a promising direction for assisting with diagnosis, or with adapting interventions to groups of patients [19]. Yarkoni and Westfall have proposed that "increased focus on prediction, rather than explanation, can ultimately lead us to greater understanding of behavior" [35], suggesting that computational tools based on machine learning and pattern recognition approaches can bring new perspectives to behavioral and clinical research.

In the last years, several scoping reviews have been published to investigate the role of machine learning in mental healthcare. Shatte et al. identified three hundred papers in 2019, on detection and diagnosis; prognosis, treatment and support; public health, and research and clinical administration [32]. Only eight of these papers were on psychotherapy, and most works focused on depression, schizophrenia, and Alzheimer's disease. In 2021, Aafjes-van Doorn et al. published a scoping review focusing on psychoanalysis, and found over fifty papers in just this subfield using machine learning [1]. Clearly, the interest in the computational tools is growing, as the capabilities of the tools are growing rapidly.

6 ACKNOWLEDGEMENTS

I would like to thank my students and colleagues, who explored these issues with me over the years. I would also like to thank all the patients and their families who have donated valuable data for the advancement of research.

REFERENCES

[1] Katie Aafjes-van Doorn, Céline Kamsteeg, Jordan Bate, and Marc Aafjes. 2021. A scoping review of machine learning in psychotherapy research. *Psychotherapy Research* 31, 1 (2021), 92–116.
[2] Thomas M Achenbach and C Edelbrock. 1991. Child behavior checklist. *Burlington (Vt)* 7 (1991), 371–392.
[3] Eda Aydın Oktay, Koray Balcı, and Albert Ali Salah. 2015. Automatic assessment of dimensional affective content in Turkish multi-party chat messages. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*. 19–24.
[4] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. 2018. The Turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 1–6.
[5] Jeffrey F Cohn. 2010. Advances in behavioral science using automated facial image analysis and synthesis [social sciences]. *IEEE Signal processing magazine* 27, 6 (2010), 128–133.
[6] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. 2017. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics* 22, 2 (2017), 525–536.
[7] Metehan Doyran, Arjan Schimmel, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander CJ Bakkes, Heysem Kaya, Ronald Poppe, and Albert Ali Salah. 2021. MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces* (2021), 1–19.
[8] Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, Sibel Halfon, and Albert Ali Salah. 2019. Video and text-based affect analysis of children in play therapy. In *ACM International Conference on Multimodal Interaction*. 26–34.
[9] Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio CS Jacques, Meysam Madadi, et al. 2021. Modeling, Recognizing, and Explaining Apparent Personality from Videos. *IEEE Transactions on Affective Computing* (2021).
[10] Binnur Görer, Albert Ali Salah, and H Levent Akın. 2017. An autonomous robotic exercise tutor for elderly people. *Autonomous Robots* 41, 3 (2017), 657–678.
[11] Sibel Halfon. 2021. Psychodynamic technique and therapeutic alliance in prediction of outcome in psychodynamic child psychotherapy. *Journal of Consulting and Clinical Psychology* 89, 2 (2021), 96.
[12] Sibel Halfon, Metehan Doyran, Batikan Turkmen, Eda Aydın Oktay, and Albert Ali Salah. 2020. Multimodal Affect Analysis of Psychodynamic Play Therapy. *Psychotherapy Research* (2020).
[13] Sibel Halfon, Eda Aydın Oktay, and Albert Ali Salah. 2016. Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis. In *International workshop on human behavior understanding*. Springer, 15–34.
[14] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* 65 (2017), 66–75.
[15] Heysem Kaya and Albert Ali Salah. 2014. Eyes whisper depression: A CCA based multimodal approach. In *Proceedings of the 22nd ACM international conference on Multimedia*. 961–964.
[16] Heysem Kaya, Albert Ali Salah, Alexey Karpov, Olga Frolova, Aleksey Grigorev, and Elena Lyakso. 2017. Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language* 46 (2017), 268–283.
[17] Paulina F Kernberg, Saralea E Chazan, and Lina Normandin. 1998. The children's play therapy instrument (CPTI): description, development, and reliability studies. *The Journal of psychotherapy practice and research* 7, 3 (1998), 196–207.
[18] David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science* 369, 6507 (2020), 1060–1062.
[19] Yves Lecrubier. 2008. Refinement of diagnosis and disease classification in psychiatry. *European archives of psychiatry and clinical neuroscience* 258, 1 (2008), 6–11.
[20] Bruno Lepri, Albert Ali Salah, Fabio Pianesi, and Alex Sandy Pentland. 2011. Human behavior understanding for inducing behavioral change: Social and theoretical aspects. In *Int. Joint Conf. on Ambient Intelligence*. Springer, 252–263.
[21] Stuart A Montgomery and Marie Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry* 134, 4 (1979), 382–389.
[22] Alejandro Moreno, Robby van Delden, Ronald Poppe, and Dennis Reidsma. 2013. Socially aware interactive playgrounds. *IEEE pervasive computing* 12, 3 (2013), 40–47.
[23] Cosmin Munteanu and Albert Ali Salah. 2017. Multimodal technologies for seniors: challenges and opportunities. In *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations-Volume 1*. 319–362.
[24] Anastasia Pampouchidou, Panagiotis G Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. 2017. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing* 10, 4 (2017), 445–470.
[25] Ronald Poppe. 2017. *Automatic Analysis of Bodily Social Signals*. Cambridge University Press, 155–167. https://doi.org/10.1017/9781316676202.012
[26] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*. 3–13.
[27] Albert Ali Salah, Jeffrey Cohn, Ronald Poppe, and Heysem Kaya. 2021. Computational Approaches to Behavioral and Clinical Science. *Submitted for publication* (2021).
[28] Albert Ali Salah and Theo Gevers. 2011. *Computer analysis of human behavior*. Springer.
[29] Albert Ali Salah, Theo Gevers, Nicu Sebe, and Alessandro Vinciarelli. 2010. Challenges of human behavior understanding. In *International Workshop on Human Behavior Understanding*. Springer, 1–12.
[30] Albert Ali Salah, Ben J. A. Kröse, and Diane J. Cook. 2015. Behavior Analysis for Elderly. In *Human Behavior Understanding*. Springer Int., Cham, 1–10.
[31] Albert Ali Salah, Nicu Sebe, and Theo Gevers. 2011. Communication and automatic interpretation of affect from facial expressions. In *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*. IGI Global, 157–183.
[32] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine* 49, 9 (2019), 1426–1448.
[33] Gizem Soğancıoğlu, Oxana Verkholyak, Heysem Kaya, Dmitrii Fedotov, Tobias Cadée, Albert Ali Salah, and Alexey Karpov. 2020. Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition. In *Proc. Interspeech 2020*. 2097–2101. https://doi.org/10.21437/Interspeech.2020-3160
[34] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.
[35] Tal Yarkoni and Jacob Westfall. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science* 12, 6 (2017), 1100–1122. https://doi.org/10.1177/1745691617693393
[36] Robert C Young, Jeffery T Biggs, Veronika E Ziegler, and Dolores A Meyer. 1978. A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry* 133, 5 (1978), 429–435.