Affect-Aware Machine Learning Models for Deception Detection

Leena Mathur
Department of Computer Science
Viterbi School of Engineering
University of Southern California
lmathur@usc.edu

Abstract

Automated deception detection systems can enhance societal well-being by helping humans detect deceivers and support people in high-stakes situations across health, social work, and legal domains. Existing computational approaches for detecting deception have not leveraged dimensional representations of affect, specifically *valence* and *arousal*, expressed during communication. My research presents a novel analysis of the potential for including affect in machine learning models for detecting deception. My work informs and motivates the development of *affect-aware machine learning approaches* for modeling deception and other social behaviors during human interactions in-the-wild. This research, independently defined and conducted by me, is from work-in-progress towards my undergraduate thesis in the Department of Computer Science at the University of Southern California.

Introduction

Advances in the fields of *social signal processing* and *multimodal machine learning* are enabling the development of automated systems and human-machine interfaces that can detect and recognize human behaviors, including the social behavior of *deception*. Deception involves the intentional communication of false or misleading information. To create a more healthy and secure society, psychologists, social workers, governments, and law enforcement groups have fostered a growing interest in detecting deception in high-stakes situations (e.g., helping social workers recognize whether clients are masking negative emotions or experiences, helping judges assess courtroom testimonies of children coerced to lie) (Burzo et al. 2018). Human deception detection abilities have been established as around chance level, motivating the development of computational approaches and automated systems that can assist humans in this task.

Background and Hypothesis

Detecting deception in videos is a current focus of the deception research community, because video-based approaches are *non-invasive* and preferable to *invasive* approaches (e.g., polygraphs) that attach to the body to collect behavioral cues indicative of deceptive communication (Burzo et al. 2018). Current machine learning approaches for video-based deception detection have exploited cues from human *visual* (e.g., head movements), *vocal* (e.g., pitch), *verbal* (e.g., word choice), and *physiological* (e.g., thermal images) modalities (Burzo et al. 2018).

Prior machine learning approaches for video-based deception detection have not leveraged representations of continuous affect, neurophysiological states (e.g., ”pleasure”) that can be components of emotions or moods. Affect can be modeled along two dimensions: how pleasant or unpleasant each state is (valence) and how passive or active each state is (arousal) (Russell 1980). Psychology research on deception has theorized that deceivers in high-stakes situations will exhibit affective states with lower valence and higher arousal, compared to truthful speakers, expressed through patterns in involuntary nonverbal cues (Ekman and Friesen 1969; Zuckerman, DePaulo, and Rosenthal 1981). Motivated by these psychology insights, *this research hypothesized that temporal patterns in representations of facial valence and facial arousal could be effectively leveraged to detect deception in videos of real-world, high-stakes situations.*

Methodology

A video dataset was used of people speaking truthfully or deceptively in real-world courtroom situations (Pérez-Rosas et al. 2015). This dataset is the current benchmark for deception detection in high-stakes situations. There were 108 videos used (53 truthful videos, 55 deceptive videos, 47 people of diverse race and gender) with an average video length of 28 seconds. This dataset's creators assigned ground truth for ”truthful” and ”deceptive” labels of videos by verifying testimonies through police investigations.

Features were extracted from facial affect, visual, vocal, and verbal modalities of communication in videos. A state-of-the-art deep neural network, trained on the Aff-Wild dataset (Kollias et al. 2018), was implemented to extract representations of facial valence and facial arousal (continuous values between -1 and 1) from each speaker's facial frames. OpenFace (Baltrusaitis et al. 2018) was used to extract 31 visual features from each speaker's facial frames, representing facial action units, head pose, and eye gaze cues. OpenSMILE (Eyben et al. 2013) was used to extract 65 vocal features from each speaker's audio frames, representing vocal frequency, energy, cepstral, loudness, and voicing information. LIWC (Kahn et al. 2007) was used to extract 93 verbal features capturing psycholinguistic information (e.g., attentional focus, cognitive mechanisms) from the transcript of each video. A fixed-length feature vector representing each feature's temporal characteristics along each video was computed with TsFresh (detailed feature descriptions in the Efficient Parameters class) (Christ et al. 2018).

Deception detection was formulated as a binary classification task to classify videos as truthful or deceptive. Experiments were conducted with *unimodal* and *multimodal* Support Vector Machine-based approaches for classifying deception. Unimodal models were trained on the respective features of the 4 modalities. Multimodal models were trained on each of the 11 multimodal combinations of the 4 modalities. Multimodal approaches included *early fusion*, *non-generative ensembles* (voting, stacking, hybrid fusion), and *generative ensembles* (bagging and boosting). All experiments were conducted with 5-fold stratified cross-validation, repeated 10 times, and split across 47 identities (the same person was never in both the training and testing folds of any experiment). Consistent with prior deception detection studies with this dataset, the primary metric used to evaluate models was average AUC computed across folds.

Key Results and Discussion

Classification results are visualized in **Fig. 1**. Unimodal models trained on facial affect features, alone, achieved an AUC of 80%. Facial affect contributed towards the best multimodal approach, which obtained an AUC of 91% and accuracy of 84% through adaptive boosting (AdaBoost) across facial affect, visual, and vocal modalities. The 91% AUC achieved by our approach was higher than the AUC of the best-performing automated approach on this dataset (88% AUC) that did not use facial affect, but also used an SVM with interpretable visual, vocal, and verbal features (Wu et al. 2018). *These results demonstrate the discriminative power of facial affect as a feature set in multimodal machine learning models for automated deception detection.* Additional results and analyses from my experiments are detailed

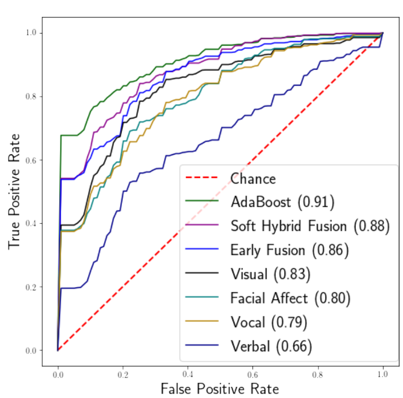

Figure 1: ROC curves for unimodal models and the best multimodal models from early fusion, non-generative ensemble (soft hybrid fusion), and generative ensemble (AdaBoost)

in a paper published at ACM's International Conference on Multimodal Interaction (Mathur and Matarić 2020).

Conclusion

This research introduced facial affect as a novel, discriminative feature set for automated deception detection in high-stakes situations. These findings provide a proof-of-concept and motivation for future research towards developing affect-aware machine learning models for detecting deception and other social behaviors during human interactions in unconstrained, real-world situations.

Acknowledgements

I would like to thank Professor Maja Matarić for advising this research effort. This research was supported by the USC Provost's Undergraduate Research Fellowship.

References

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66.

Burzo, M.; Abouelenien, M.; Perez-Rosas, V.; and Mihalcea, R. 2018. *Multimodal Deception Detection*, 419–453. ISBN 9781970001716.

Christ, M.; Braun, N.; Neuffer, J.; and Kempa-Liehr, A. 2018. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests doi:10.1016/j.neucom.2018.03.067.

Ekman, P.; and Friesen, W. 1969. *Nonverbal Leakage and Clues to Deception*. Defense Technical Information Center.

Eyben, F.; Weninger, F.; Gross, F.; and Schuller, B. 2013. Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. ISBN 9781450324045. doi:10.1145/2502081.2502224.

Kahn, J.; Tobin, R.; Massey, A.; and Anderson, J. 2007. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American journal of psychology* 120: 263–86. doi:10.2307/20445398.

Kollias, D.; Tzirakis, P.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; Schuller, B. W.; Kotsia, I.; and Zafeiriou, S. 2018. Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *CoRR* abs/1804.10938.

Mathur, L.; and Matarić, M. J. 2020. Introducing Representations of Facial Affect in Automated Multimodal Deception Detection. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, 305–314. doi:10.1145/3382507.3418864.

Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; and Burzo, M. 2015. Deception Detection Using Real-Life Trial Data. In *2015 ACM on International Conference on Multimodal Interaction*. doi:10.1145/2818346.2820758.

Russell, J. 1980. A Circumplex Model of Affect. *J Pers Soc Psychol* 39: 1161–1178. doi:10.1037/h0077714.

Wu, Z.; Singh, B.; Davis, L. S.; and Subrahmanian, V. S. 2018. Deception Detection in Videos. In *AAAI Conference on Artificial Intelligence*.

Zuckerman, M.; DePaulo, B. M.; and Rosenthal, R. 1981. *Verbal and Nonverbal Communication of Deception*, 1–59. doi:10.1016/s0065-2601(08)60369-x.