IDENTIFYING TRUTHFUL LANGUAGE IN CHILD INTERVIEWS

Victor Ardulov[1], *Zane Durante*[1], *Shanna Williams*[2], *Thomas Lyon,*[2] *Shrikanth Narayanan*[1]

Signal Analysis and Interpretation Lab[1]
Gould School of Law [2]
University of Southern California
Los Angeles, CA

ABSTRACT

When a child is suspected to be the victim or sole witness of a crime, the manner in which information is gathered from the child becomes critical. A child forensic interview is the guided conversation that a legal expert conducts to elicit reliable information from a child. To help substantiate child testimony, it is important to discern characteristics of truthful and deceptive behavior in these interviews. The work presented uses various machine learning algorithms to identify differences in the speech of children when they are lying or being truthful, particularly when they have been asked by a confederate to deceive an interviewer. Results show that vocabulary and psycho-linguistic norms of a child's language use, in response to directed questions, provide substantial information to outperform human adults in detecting truthful statements.

***Index Terms*—** Behavioral Signal Processing, Child Forensic Interview, Deception Detection

1. INTRODUCTION

A child forensic interview is administered in order to elicit testimony when children are suspected of being the victim of, or witness to a crime [1]. In particular, as children are often the victims of crimes committed by their guardians, children may feel conflicted to disclose information or even be coached to deny wrong-doing. Additionally, recalling traumatic events can induce stress, leading to false denials of crimes which exonerate criminals [2]. Similarly, false allegations must also be avoided because of the profound legal consequences of child testimony for the accused. As a result, understanding and detecting child deception is critical in the context of legal proceedings.

Identifying deception in children is challenging as their nascent capacity to communicate confounds the process. Consequently, adults are found to be poor judges of deception in children and only discern truthfulness from deception with an average accuracy of 54% [3]. Due to the poor reliability of human judgments of deception in children, various methods have been developed to automatically detect deception. Previous research in automated deception detection has primarily focused on the qualitative and quantitative differences between truthful and deceptive statements in courtroom and forensic testimony with adults [4, 5]. These studies have shown success in using many modalities including language in predicting when a defendant's testimony of innocence is consistent with the verdict ruled against them. However, an analysis of deception detection in children demonstrated that verbal markers discerning truthfulness from deception can be obscured with parent coaching [6, 7]. Machine learning models have been evaluated using measures of syntactical complexity from real-life court testimony, including features such as average word complexity and utterance word length [8]. However, automated detection of child deception remains a largely unexplored area.

The presented work builds on existing approaches, analyzing how the vocabulary and emotional content of the language used by children indicates truthfulness. The results and analysis of multiple models and input configurations are evaluated to gain further insight into child truthfulness. A further analysis points to specific psycho-linguistic features that models correlate with truthfulness.

2. BACKGROUND

The physical and psychological immaturity of children makes them vulnerable to abuse, affecting their mental health into adolescence and adulthood [9, 10, 11]. Since children are most often victims or witnesses to crimes committed by their legal guardians or caretakers, eliciting accurate and truthful testimony can remove them from dangerous living environments. However, this same context can create a conflicting desire to testify for the child [12, 13]. This challenge is further compounded because the same elements of a child's cognitive and language development that contribute to their vulnerability also influence their ability to consistently recount incidents in courtroom settings. This makes their testimony subject to coercion and intimidation, often leading to its dismissal in court [14].

Child forensic interview (CFI) protocols have been developed to protect children from coercion and intimidation while obtaining their testimony [1, 15]. Previous work has evaluated psychological effects of interviewer language and CFI prompts on child responses in court proceedings [16, 17]. However, it is inappropriate and sometimes impossible to assign ground truth labels to assess the truthfulness of a child's statements, limiting the work of computational models to proxied values of verbal productivity [18, 19]. To overcome these shortcomings, this work utilizes a set of simulated interviews in which interviewers practice the CFI protocol in order to elicit a child's disclosure to a toy breaking [20].

A meta-analysis of detection of deception by adults in these same settings demonstrated that adults typically have an accuracy of 0.54 for correctly discerning when a child is lying or when they are telling the truth [3]. Previous work on the automated detection of deception has been largely confined to adult subjects, utilizing multi-modal streams of features including video, audio, and text [5, 21]. Meanwhile, work on automatic deception detection in children has focused on syntactic analysis of child speech, evaluating features such as average word complexity and utterance word length [8]. The presented research builds on this work by looking at the predictive power of specific vocabulary and psycho-linguistics.

This work compares the effectiveness of various machine learning and natural language processing techniques to predict deception in these transcribed interviews. Linguistic features from transcripts in different phases of CFI across individuals are used to generate predictions of truthful disclosure.

3. METHODS

3.1. Data

The data consist of 217 transcribed interactions between a unique child and 2 adult experimenters: a *confederate*, and an *interviewer*. These simulated interviews were collected by legal psychologists trained in the CFI protocol. The children included in this study are known to be elementary school aged, but specific demographic data is not available for the children.

Each interaction consists of three phases: confederate play, rapport building, and recall, occurring in that order. An interaction begins with a confederate engaging the child with some toys. Prior to the interaction, the confederate is informed about whether or not any of the toys have been rigged to break during play. The event of a toy breaking is called a transgression and henceforth, T will denote transgression and T' denotes no transgression occurring. Towards the end of playing with the toys, the confederate will inform the child that an interviewer will enter the room to ask them some questions. Given T, the confederate will ask the child to promise not to disclose to the interviewer that the toy broke. The confederate then exits the room, and the interviewer enters. Upon entering the room the interviewer is unaware as to whether or not a transgression has transpired. The interviewer begins by following a modified CFI protocol for *rapport building*, adhering to a semi-structured script that does not discuss the broken toys. Once the interviewer has completed rapport building, they will transition to the final stage, *recall*. During recall, the child is asked to name every toy and narrate everything that happened with each toy. The interviewer will only repeat a question if the child indicates they did not understand the question.

The interaction is then transcribed and annotated for disclosure labels. Specifically, the questions are coded for the child's willing admission to transgression. Disclosure is denoted by D for disclosure and D' for no disclosure.

	Disclosure	
Transgression	D	D'
T	44(20.3%)	121(55.8%)
T'	2(< 1%)	50(23.0%)

Table 1. Number of transcripts between different transgression and disclosure conditions

Transcripts are categorized into four groups: $(D|T)$, $(D|T')$, $(D'|T)$, $(D'|T')$ with respective distributions reflected in Table 1. These categories were then grouped into two broader classes: *truth* $(D|T) \cup (D'|T')$ and *deception* $(D'|T) \cup (D|T')$.

The transcripts were pre-processed to remove stop words and punctuation and converted into vectors representing term frequency-inverse document frequency (tf-idf) for each session. The transcripts were evaluated using unigrams and bigrams, separately and together. Furthermore, the psycholinguistic norms for valence, arousal, pleasantness, and age of acquisition were summed over the n-grams in the transcripts in order to capture the associated affective constructs of the language used as defined by EmotiWord [22].

3.2. Baseline

According to a meta-analysis of adult detection of child truth telling, it was reported that, across different modalities and conditions, on average adult humans had an accuracy of 0.54 [3]. In order to compare the models' predictive performance against adults, 10000 users were simulated in order to establish a distribution over the expected human performance. Each simulated user would "predict" the label for a transcript and a probability of 0.54 of correctly assigning the label. Each user's F1 score was then recorded. Each simulated user and simulated prediction was considered independent.

3.3. Models

Given the limited amount of data, decision trees (DT) [23], random forests (RF) [24], and a 2 layer neural network (NN) were the three models taken into consideration. The neural network consisted of a ReLU activated hidden layer connected to a soft-max activated output layer. This specific neu-

		Rapport			Recall			Both		
Data	Model	unigram	bigram	both	unigram	bigram	both	unigram	bigram	both
CO	NN	0.5162	0.3410	0.4078	0.4300	0.2310	0.4276	0.5278	0.1292	0.4486
	DT	0.4250	0.2118	0.4592	**0.5702**	0.2310	0.4746	0.5518	0.3634	**0.5718**
	RF	0.3498	0.2518	0.3494	**0.6226**	0.1084	0.5628	**0.5786**	0.2518	0.5194
C+I	NN	0.4604	0.4636	0.4360	0.5280	0.4926	0.3710	0.4608	0.3602	0.5284
	DT	0.3754	0.4976	0.4430	0.4744	0.2376	0.4368	0.5340	0.4784	0.5518
	RF	0.3580	0.3892	0.4020	**0.6196**	0.1292	**0.6090**	0.5474	0.2326	0.5630

Table 2. Resulting F1 for Task 1 using only tf-idf

		Rapport			Recall			Both		
Data	Model	unigram	bigram	both	unigram	bigram	both	unigram	bigram	both
CO	NN	0.4492	0.2844	0.3844	0.3304	**0.6170**	0.4700	**0.5882**	0.2310	0.3160
	DT	0.3476	0.3514	0.4346	0.4372	**0.6240**	0.3908	0.3410	0.5240	0.3796
	RF	0.3702	0.3770	0.3350	0.5630	0.5400	0.5392	0.5590	0.3680	0.5322
C+I	NN	0.3014	0.3742	0.4528	0.3718	**0.6198**	0.5686	0.4806	0.3692	0.4342
	DT	0.4738	0.4738	0.5034	0.4864	**0.6186**	0.4744	0.3406	0.3556	0.5464
	RF	0.3754	0.4210	0.3414	0.5568	0.5526	**0.6190**	0.5400	0.4268	0.5458

Table 3. Resulting F1 for Task 1 using both tf-idf and psycho-linguistic norms

ral network architecture was used to compare against previous work in deception detection in adults [5].

3.4. Evaluation Tasks

Two tasks were considered for the evaluation of the models: *Task 1* used the deception and truth classes and evaluated the models' ability to correctly discern between the two. Due to the limited size of $(D|T')$, in *Task 2* the models were trained on all categories, but were only evaluated on the subset of deception and truth in the D' condition.

To overcome auspicious data splits, we report the models' average F1 score across 5-fold cross validation splits. The experiments evaluated the performance across data representations, namely unigram, bigram, and psycho-linguistic norms as part of the model input. Experiments also evaluated the effect of the interviewers' language use by considering child and interviewer (C+I) and child only (CO) inputs. Finally, the models were assessed on the different sections of the interview (Rapport, Recall, Both). This resulted in 108 combinations of input data representation and model type per task.

4. RESULTS

Simulating separate baseline F1 scores for the different tasks yielded an average of 0.5073 for Task 1, where both disclosure conditions were considered, and 0.4069 for Task 2, when considering only the non-disclosure condition. Tables 2 - 5 show the average cross validation F1 score for models. The highlighted values represent models which are considered significant improvements over the baseline for producing F1 score greater than the 95th percentile of the distribution of simulated F1 scores. Significant improvement baselines were 0.5702 and 0.4812 for Tasks 1 and 2 respectively.

For Task 1, 12 experiments produced average F1 scores performing significantly better than their respective baseline, while only 3 experiments performed significantly better for Task 2. Generally, when not considering psycho-linguistic norms, bigrams performed worse than unigrams. Conversely when psycho-linguistic norms were considered, bigram models typically outperformed unigram models when the Recall stage was being included. This trend is captured in both Task 1 and Task 2 and accounts for the split that contains the majority of models that exceed the baseline. The inclusion of interviewer language largely hindered the performance of the models or added insignificant improvements.

4.1. Psycho-Linguistic Analysis

Due to the general increase in performance when psycho-linguistic norms were included, the predictions of the decision tree model using bigrams and psycho-linguistic norms of the CO-Recall condition were evaluated for correlation of the input psycho-linguistic norms.

Table 6 reflects that high valence, arousal, and age of acquisition are all correlated with a higher probability assignment to truth. Conversely, pleasantness had a negative correlation with predictions of truthfulness.

5. DISCUSSION

The relative drop in F1 score between the two different tasks suggests that the models are learning to identify disclosure from non-disclosure. This is motivated by the fact that in Task

		Rapport			Recall			Both		
Data	Model	unigram	bigram	both	unigram	bigram	both	unigram	bigram	both
CO	NN	0.3168	0.1900	0.2546	0.3548	0.3246	0.3058	0.2974	0.1980	0.2504
	DT	0.3670	0.2484	0.3448	0.3794	0.4484	0.3024	0.3206	0.3524	0.2968
	RF	0.3424	0.2564	0.3658	0.2566	0.0762	0.2556	0.2598	0.3782	0.2730
C+I	NN	0.3540	0.1444	0.3550	0.3002	0.3702	0.3856	0.4666	0.1900	0.2720
	DT	0.3408	0.3702	0.2962	0.2006	0.2742	0.4136	0.3780	0.1900	0.3874
	RF	0.3430	0.2564	0.3068	0.3744	0.3702	0.3452	0.2694	0.0762	0.3228

Table 4. Resulting F1 for Task 2 using only tf-idf

		Rapport			Recall			Both		
Data	Model	unigram	bigram	both	unigram	bigram	both	unigram	bigram	both
CO	NN	0.4602	0.2600	0.4352	0.4564	0.4588	0.3166	0.4560	0.2772	0.3442
	DT	0.4248	0.2252	**0.5556**	0.3264	**0.5322**	0.3966	**0.4928**	0.1490	0.4470
	RF	0.4402	0.3486	0.2810	0.2354	0.3542	0.3048	0.2718	0.3708	0.2790
C+I	NN	0.3686	0.2562	0.3290	0.4566	0.4628	0.2524	0.2740	0.3186	0.1918
	DT	0.3494	0.3514	0.3170	0.4328	0.4654	0.1402	0.4338	0.1170	0.3724
	RF	0.2890	0.3644	0.3150	0.3386	0.3652	0.4286	0.3494	0.3084	0.2150

Table 5. Resulting F1 for Task 2 using both tf-idf and psycho-linguistic norms

Psycho-linguistic Norm	Correlation
Valence	0.1768
Arousal	0.2081
Pleasantness	-0.2554
Age of Acquisition	0.2429

Table 6. Pearson Correlation between psycho-linguistic norms and model output *truth* on Task 1 for decision trees during free recall. All values have $p < 0.05$.

2 there are no models that perform better than the baseline when psycho-linguistic norms are not considered. Furthermore the results suggest that much of the signal in identifying truth telling comes from the psycho-linguistic norms. In particular, the results of the cross-correlational analysis suggests that while valence, arousal, and age of acquisition positively correlate to truthfulness, pleasantness is negatively correlated. This may imply that children use words that seem more pleasant to mask their deception. The relationship with age of acquisition also suggests that more complex or descriptive language is used when describing the truth. Or, if age of acquisition is used as a proxy for the child's cognitive development, this correlation implies that children with more linguistic development are more likely to tell the truth. An alternative explanation may be that children revert to simpler language when they are trying to lie, which may garner insights into the cognitive load of deception in children.

Generally, recall had more significant lexical information as to whether the child is telling the truth when compared to rapport building. Accordingly, the models tended to perform better when analyzing the language pertinent to disclosure. This observation may indicate that specific linguistic and psycho-linguistic shifts occur in the child's language use when they are asked to narrate a transgression.

Ultimately, the results indicate that children are very effective at deception when it involves concealment as opposed to out-right fabrication. However, the vocabulary used by children can provide meaningful insights as to whether they are being truthful or deceptive. Moreover, the affective information carried by those words seem to be a more robust feature to observe.

6. CONCLUSION

This project evaluated an initial collection of data and model pairings and their relative performances to detect deception. The fact that fairly low-resource and low complexity models can outperform human adults significantly across a single modality indicates the power of signal processing techniques to further our insight into the ways that children might behave during interviews.

In the future, we should continue to look for more subtle means of detecting deception. Interlocutor dynamics over time across multiple modalities, including semantic information of children's vocabulary, may prove to extend the findings of this work.

7. ACKNOWLEDGMENTS

The authors would like to extend their gratitude to and acknowledge the participants of the experiments and those who transcribed and coded the interactions.

8. REFERENCES

[1] T. D. Lyon, “Ten step investigative interview,” *Los Angeles, CA: Author*, 2005.

[2] T. D. Lyon, “False denials: Overcoming methodological biases in abuse disclosure research,” in *Child sexual abuse*, pp. 51–72. Psychology Press, 2007.

[3] J. Gongola, N. Scurich, and J. A. Quas, “Detecting deception in children: A meta-analysis,” *Law and Human Behavior*, vol. 41, no. 1, pp. 44–54, 2017.

[4] S. L. Sporer, “Reality monitoring and detection of deception,” *The detection of deception in forensic contexts*, pp. 64–102, 2004.

[5] R. Mihalcea and C. Strapparava, “The lie detector: Explorations in the automatic recognition of deceptive language,” in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Stroudsburg, PA, USA, 2009, ACLShort ’09, pp. 309–312, Association for Computational Linguistics.

[6] C. Saykaly, V. Talwar, R. C. L. Lindsay, and K. Lee, “The influence of multiple interviews on the verbal markers of children’s deception,” *Law and human behavior*, vol. 37, pp. 187–96, 06 2013.

[7] V. Talwar, K. Hubbard, C. Saykaly, K. Lee, R. C.L. Lindsay, and N. Bala, “Does parental coaching affect children’s false reports? Comparing verbal markers of deception,” *Behavioral Sciences and the Law*, vol. 36, no. 1, pp. 84–97, 2018.

[8] M. Yancheva and F. Rudzicz, “Automatic detection of deception in child-produced speech using syntactic complexity features,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, vol. 1, pp. 944–953.

[9] T. Hillberg, C. Hamilton-Giachritsis, and L. Dixon, “Review of meta-analyses on the association between child sexual abuse and adult mental health difficulties: A systematic approach,” *Trauma, Violence, & Abuse*, vol. 12, no. 1, pp. 38–49, 2011.

[10] D. M. Fergusson, L. J. Horwood, and M. T. Lynskey, “Childhood sexual abuse and psychiatric disorder in young adulthood: Ii. psychiatric outcomes of childhood sexual abuse,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 35, no. 10, pp. 1365–1374, 1996.

[11] J. B. Kaplow, E. Hall, K. C. Koenen, K. A. Dodge, and L. Amaya-Jackson, “Dissociation predicts later attention problems in sexually abused children,” *Child Abuse & Neglect*, vol. 32, no. 2, pp. 261–275, 2008.

[12] L. Radford, S. Corral, C. Bradley, H. Fisher, C. Bassett, N. Howat, and S. Collishaw, “Child abuse and neglect in the uk today,” *London: NSPCC*, 2011.

[13] M. E. Lamb, K. J. Sternberg, Y. Orbach, P. W. Esplin, H. Stewart, and S. Mitchell, “Age differences in young children’s responses to open-ended invitations in the course of forensic interviews.,” *Journal of consulting and clinical psychology*, vol. 71, no. 5, pp. 926, 2003.

[14] T. D. Lyon, S. N. Stolzenberg, and K. McWilliams, “Wrongful acquittals of sexual abuse,” *Journal of interpersonal violence*, vol. 32, no. 6, pp. 805–825, 2017.

[15] M. E. Lamb, Y. Orbach, I. Hershkowitz, P. W. Esplin, and D. Horowitz, “A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol,” *Child abuse & neglect*, vol. 31, no. 11-12, pp. 1201–1231, 2007.

[16] M. E. Lamb and A. Fauchier, “The effects of question type on self-contradictions by children in the course of forensic interviews,” *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 15, no. 5, pp. 483–491, 2001.

[17] M. S. Brady, D. A. Poole, A. R. Warren, and H. R. Jones, “Young children’s responses to yes-no questions: Patterns and problems,” *Applied Developmental Science*, vol. 3, no. 1, pp. 47–57, 1999.

[18] V. Ardulov, M. Kumar, S. Williams, T. D. Lyon, and S. Narayanan, “Measuring conversational productivity in child forensic interviews,” *arXiv e-prints*, p. arXiv:1806.03357, Jun 2018.

[19] V. Ardulov, M. Mendlen, M. Kumar, N. Anand, S. Williams, T. D. Lyon, and S. Narayanan, “Multimodal interaction modeling of child forensic interviewing,” in *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*, 2018, pp. 179–185.

[20] T. D. Lyon, L. C. Malloy, J. A. Quas, and V. A. Talwar, “Coaching, truth induction, and young maltreated children’s false allegations and false denials,” *Child Development*, vol. 79, no. 4, pp. 914–929, 2008.

[21] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.

[22] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Emotiword: Affective lexicon creation with application to interaction and multimedia data,” in *MUSCLE*, 2011.

[23] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees. wadsworth int,” *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[24] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.