AI Lie Detection: Is the Hype Justified?

Stanford CS224N Custom Project

Jack Ryan
Department of Computer Science
Stanford University
ryanjack@stanford.edu

Abstract

Multimodal machine learning for deception detection is a growing field, boasting results up to 100% classification accuracy on truthful vs. deceptive statements (Constâncio et al., 2023). Considering the difficulty of true, generalizable deception detection, this is suspicious. This project tested the replicability of one of these results, Gogate et al. (2017), and closely investigated whether the result was only possible due to methodological issues. Evidence has been found suggesting between 23-26% accuracy boosts when poor methodological choices are made, between a baseline and the replicated model. Though the primary source of improper performance gains was from changes in the data split random seed, with performance varying up to 37% among only 10 random seeds, using untrained models (on top of word embeddings). This result suggests that the data samples are not independent, since random, independent guessing on each sample is unlikely to lead to performance variation this large, though this conclusion needs further examination. If true, this would be cause for skepticism about all results involving the Real-Life Trial Dataset (Pérez-Rosas et al., 2015), the most common dataset used for deception detection with machine learning (Constâncio et al., 2023), since a lack of independence is approximately equivalent to using a smaller, independent dataset, and the dataset is already small with only 121 samples.

1 Key Information to include

- Mentor: Tathagat Verma

2 Introduction

Polygraph tests are not currently permissible as evidence in court in most US states. This is because research has shown that their results aren't generally reliable, and they come with technical caveats. Due to this, courts have judged that their scientific nature may cause juries to trust them more than is warranted, and so they are not permissible evidence (Cavoukian and Heslegrave (1980)). On the other hand, the past decade has seen the rise of deep learning, and subsequently, a rise in machine learning deception detection. For instance, in Constâncio et al. (2023), 81 research papers on deception detection using machine learning were analyzed, and they found results ranged from 51% to 100% accuracy, with 19 works reporting above 90% accuracy.

But if these results are legitimate, why hasn't there been as much discussion about these systems as there has been about the polygraph? Highly performant AI deception detection systems would be significantly change the world, and courts would be ruling on their usage as evidence in trials. But none of this appears to be happening.

These results are suspicious for other reasons too. Truly generalizable deception detection intuitively should be very difficult, if possible at all. Additionally, many of these results use the Real-Life Trial Deception Detection Dataset which only consists of only 121 samples (Pérez-Rosas et al., 2015).

And no research testing the validity and reliability of these results was found (though it may exist). This sort of research could be significant, considering in the case of the polygraph that this type of research has influenced several court policies.

For these reasons, this paper aims to test the validity and reliability of these deception detection results by replicating one of these papers, specifically Gogate et al. (2017), which uses the Real-Life Trial Deception Detection Dataset, and which appears to have several methodological flaws. In addition to replication, methodological analyses were done to test whether improper methodological choices can improve performance significantly in this domain. It was found that certain poor methodological choices did improve accuracy significantly, up to 23-26% for our baseline and our replicated model. Additionally, random seeds used for data splitting led to large differences in untrained model performance, with accuracies between 33.33% and 70.83% across 10 random seeds for our untrained baseline.

3 Related Work

3.1 ML Deception Detection

In Constâncio et al. (2023), a systematic review was done of the field of deception detection using machine learning. 540 distinct articles were screened, and of those, 81 were selected according the the requirements that the article had to: be about deception detection; use machine learning; clearly state the used features; report performance; and involve non-invasive methods only (e.g. no MRI). Among these papers, reported results ranged from 51% to 100% accuracy, with 19 works reporting above 90% accuracy, and two works reporting 100% accuracy. In all cases, the task was binary classification among two labels, deceptive or truthful, and the primary modalities were text, speech, and video, though other modalities were also used, such as thermal, physiological, and emotional modalities. Both unimodal and multimodal models have been used. Both traditional ML (decision trees, random forests, etc) and deep learning approaches have been used. And several datasets have been used, including multimodal and unimodal datasets of various sizes.

The most widely used dataset was the Real-Life Trial Deception Detection Dataset (Pérez-Rosas et al., 2015)), with 17 papers using it or a modified version of it. This dataset is discussed more in the Dataset section, since it's also the dataset used for replication. Out of papers using this dataset, six of them report above 90% accuracy or AUROC, and out of these six, one uses Adaboost, two use a neural network, one uses combined methods, one uses multi-view learning, and one uses logistic regression.

3.2 The Replicated Paper

Gogate et al. (2017) was one of the neural network approaches achieving above 90% accuracy on the Real-Life Trial Dataset. This paper was chosen for replication, because it reports results for a unimodal text-based classifier, and since the paper appears to have several methodological flaws.

According to the systematic review of deep learning papers for deception detection done in Constâncio et al. (2023), this paper was one of the early papers working with the Real-life Trial Deception Detection Dataset. With that context, the main contributions of the paper are:

- Introduces deep learning multimodal fusion to the problem of deception detection, utilizing a CNN architecture;
- SOTA results for: text-only deception detection (83.78%), visual-only deception detection (78.57%), and several combinations of modalities, including early vs. late fusion;
- First time use of audio cues (as far as authors know) for deception detection.

One major limitation of this paper is that they did not use a validation set, only a training set (70%) and a test set (30%). This is a major flaw, since then their hyperparameters are likely tuned to the test set, meaning the results would not generalize. Additionally, they most likely did not split the training and test sets by speaker, meaning the same speaker's clips could be in the training set and the test set. It's possible they actually did do this, but it's not mentioned when they discuss the training/test split. Additionally, they don't mention using any kind of regularization, shuffling of the training data, or fixation of random seeds, and each of these methods would be especially important to use since the dataset is small (121 clips). Also, they left out some training details (number of epochs, RMSProp parameters, and batch size), making replication more difficult. Also, the authors did very little in terms of discussing the reported results - they didn't make claims about generalization and they didn't analyze the model's failure modes.

These limitations, especially the lack of use of a validation set, make the results of this paper much less convincing, since it seems much less likely the results generalize.

3.3 Methodological Critiques

Some prior work was found in studying methodological flaws of machine learning research, primarily in medical contexts. For instance, Maleki et al. (2023) discusses three main methodological pitfalls, having to do with data augmentation, data splitting, and evaluation, and measures the effect of these pitfalls using differences in F1 score when these pitfalls are present vs. not, which is similar to the approach taken in this paper. Additionally, Varoquaux and Cheplygina (2022) had results suggesting that the error in evaluation of medical imaging models is often greater than the marginal performance improvements on Kaggle Benchmarks. The importance of proper evaluation is a theme in this research, as well as in this work.

Outside of machine learning, there is prior research which scrutinizes the validity and reliability of polygraphs for lie detection, such as Saxe et al. (1985) and Saxe and Ben-Shakhar (1999). This area of research has contributed to decisions by most US states to disallow polygraph results as evidence in court (Cavoukian and Heslegrave, 1980), suggesting that research which tests the validity of lie detection systems can have significant real-world impacts. Even so, no prior work was found that tested the validity/reliability of AI deception detection systems, though it may exist.

4 Approach

4.1 The Model

Only the unimodal text-based model from Gogate et al. (2017) was replicated, which achieved 83.78% accuracy in the paper. The model used is a CNN over concatenated 300-dimensional GloVe word embeddings, which can be visualized in Figure 1 (Pennington et al., 2014). Transcripts are padded or trimmed to 100 words, and words unrecognized by GloVe were given the same embedding as the pad token (all zeros). The architecture of the CNN is: 4 x (Conv layer, ReLU, Max pool) followed by 2 x (Linear layer, ReLU) and then 1 x (Linear layer, softmax). The Conv layers each had 15 filters, and window size 2, with padding='same'. The max pools has windows size 2. And the linear layers were (in features=1620, out features=5000), (in features=5000, out features=500), and (in features=500, out features=2) respectively. A linear-esque baseline was used over the GloVe embeddings, consisting of a broadcast linear layer across each word embedding, with out features=1, followed by a linear layer with in features=100, out features=2, followed by a softmax. Essentially, this computes the contribution of each word, then combines the contributions without non-linearities. As in the paper, RMSprop (Equations 1 and 2) and cross-entropy loss (Equation 3) were used to train both models.

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta)g_t^2 \tag{1}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{2}$$

$$H(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{3}$$

4.2 Ablation Studies

Using both the baseline and the replicated model ("Baseline" and "GogateCNN"), the main approach of this paper was to measure the difference in test set accuracy when individual, poor methodological choices are made compared to when no poor methodological choices are made. The poor methodological choices tested were: tuning hyperparameters to the test set instead the validation set; not splitting

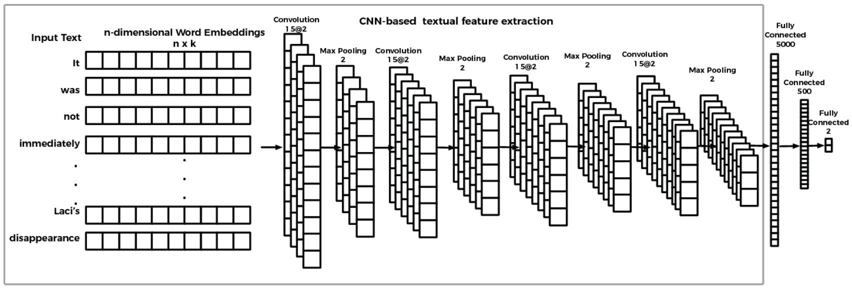

Figure 1: Unimodal textual CNN binary classifier from Gogate et al. (2017).

the data by speaker; reporting the best performance across 5 or 10 random seeds; not shuffling the training data; not using regularization; making all of these poor choices, except taking the max across random seeds; and making all of these poor choices, taking the max across 5 or 10 random seeds. Each condition was repeated 5 times with different random seeds for data splitting, except the "all good choices" and "all bad choices" conditions were repeated 10 times each, so the "max of 10 seeds" could be applied.

4.3 Random Seeds

Different random seed variables were used for different sources of randomness. Random seeds for Bayesian optimization, weight initialization, and data shuffling were fixed for all experiments. When experiments are run with different "random seeds," the random seed that changes is always only the random seed used for data splitting. The used random seeds are the same 5-10 seeds in all experiments, and their derivation is described in the Data Splitting subsection below.

4.4 Hyperparameters

Most hyperparameters were fixed across all experiments, but in order to achieve performant results, tuning was used for RMSprop parameters of learning rate, momentum, alpha, and weight decay. Except weight decay was fixed to 0 when regularization was not used. The validation set was used for tuning, except in the ablations involving tuning to the test set. In order to ensure that the same amount of optimization was applied to all scenarios, automatic hyperparameter tuning was used and redone for each ablation condition and for each random seed. In total, 2 models x (2 conditions x 10 seeds + 4 conditions x 5 seeds) = 40 separate automatic hyperparameter tuning runs were performed. Manual hyperparameter tuning was entirely avoided, even for the hyperparameters of the automatic tuning process itself. For all non-tuned hyperparameters, reasonable defaults were chosen before any experiments took place and remained unchanged throughout. Ray Tune was used for automatic tuning, specifically their Bayesian Optimization API.

4.5 Random Seed Analysis

Additionally, to examine the amount of variation coming purely from how the data is randomly split, the untrained Baseline and the untrained GogateCNN were tested on the respective validation set for random seeds 0 through 99. These seeds were only used for data splitting. No training or tuning was done, and the hyperparameters chosen were the “good methodology” choices, or reasonable defaults for the hyperparameters that were normally tuned. Due to the difficulty of finding balanced random seeds, splitting by speaker was not done.

4.6 Data Splitting

When splitting data by speaker, the set of samples for Jodi Arias and Andrea Sneiderman were always put in the training set, since the sets are of size 21 and 12, which would make test/val not diverse enough. Additionally, an algorithm was developed for ensuring train/val/test each had equal number amounts of deceptive/truthful samples when splitting the data by speaker. The algorithm sorts groups of samples, grouped by speaker, into the different sets, making it more likely the final splits are evenly deceptive/truthful. Then, using this algorithm to narrow the search space, random seeds were found which led to train/test/val always having 50%/50% deceptive/truthful samples. When splitting by speaker was not done, the same outcome was ensured using sklearn.model_selection.train_test_split. And in this case, the same random seeds were used as those used when splitting by speaker, to reduce variation between conditions.

5 Experiments

5.1 Data

The dataset used are the transcripts from the Real-Life Trial Deception Detection Dataset from Pérez-Rosas et al. (2015), consisting of 121 clips taken from court trials, labelled either deceptive or truthful, based on the outcome of the trial, as well as evidence from police investigations. Due to several spelling errors found, a manually cleaned version of the transcripts were used. This, along with punctuation stripping, reduced the number of words in the dataset which GloVe did not have an embedding for from 533 to 56, out of 1545 total unique words in the dataset.

The associated task is binary classification over the labels deceptive or truthful, given a video of a speaker (in our case, just their transcript).

5.2 Evaluation method

The evaluation metric for the ablation experiments is the difference in accuracy between the condition when a poor methodological choice is made, and when no poor choices are made. For the random seed analysis, the metrics are mean accuracy, max accuracy, and min accuracy, across random seeds.

5.3 Experimental details

Several parameters were fixed for all experiments, to reasonable defaults: test size=20%, val size=20%, eps=1e-08 (RMS prop), centered=True (RMS prop), batch size=10 (for training speed), epochs=150, number of random steps=4 (Bayesian optimization), number of samples=20 (Bayesian optimization), search space for learning rate=loguniform(1e-4, 1), search space for alpha=uniform(0.9, 0.999), search space for momentum=uniform(0.1, 0.9), and search space for weight decay=uniform(1e-4, 1e-1). Additionally, for the random seed analysis, learning rate=1e-2, alpha=0.95, momentum=0.5, and weight decay=1e-2. These valeus were chosen as the rough middle of their respective ranges. Random seeds for weight initialization, Bayesian optimization, and data shuffling were all set to 0. Random seeds for data splitting were [329, 723, 769, 977, 1230] when 5 were used and [329, 723, 769, 977, 1230, 1492, 1746, 2071, 2226, 2334] when 10 were used.

Tuning/training runs were done using Google Colab's V100. Ray Tune allowed for simultaneous utilization of the 2 provided CPUs and the GPU. All three processors were used for each sample of Bayesian optimization, rather than splitting the processors among several samples.

5.4 Results

First, the results of replication for the Baseline and the GogateCNN are presented in Table 1, where only good methodological choices were made. The average accuracy plus or minus standard deviation is shown, as well as the difference from the original result. With our setup, we were unable to achieve the original result when using proper methodology.

Model	**Accuracy**	**Difference to Original**
Baseline	$52.08\% \pm 15.17\%$	-31.70%
GogateCNN (replication)	$50.00\% \pm 0.00\%$	-33.78%
GogateCNN (original)	83.78%	N/A

Table 1: Accuracy of the replication, when only good methodological choices were made, vs. the accuracy reported in the paper, whose methodology is uncertain.

Tables 2 and 3 show the results of the ablation studies for the Baseline and the GogateCNN respectively. Mean accuracy plus or minus standard deviation is shown, along with the difference to the respective "good choices only" result found in Table 1.

Bad Methodological Choice	**Accuracy**	**Difference to Baseline**
Hyperparam-tuning to test	66.67% ± 2.64%	+14.59%
Not data-splitting by speaker	50.40% ± 5.43%	−1.68%
Using max across 5 random seeds	75.00%	+22.92%
Using max across 10 random seeds	75.00%	+22.92%
Not shuffling train data	48.33% ± 8.98%	−3.75%
Not using regularization	56.67% ± 8.58%	+4.59%
All of the above (avg)	64.40% ± 6.31%	+12.32%
All of the above (max of 5)	76.00%	+23.92%
All of the above (max of 10)	76.00%	+23.92%

Table 2: Ablation study results for Baseline model.

Bad Methodological Choice	**Accuracy**	**Difference to GogateCNN (replication)**
Hyperparam-tuning to test	50.00% ± 0.00%	0.00%
Not data-splitting by speaker	48.80% ± 1.60%	−1.20%
Using max across 5 random seeds	50.00%	0.00%
Using max across 10 random seeds	50.00%	0.00%
Not shuffling train data	49.17% ± 1.67%	−0.83%
Not using regularization	60.00% ± 12.25%	+10.00%
All of the above (avg)	62.80% ± 8.21%	+12.80%
All of the above (max of 5)	76.00%	+26.00%
All of the above (max of 10)	76.00%	+26.00%

Table 3: Ablation study results for GogateCNN model.

The results of Tables 1, 2, and 3 can also be visualized in Figures 2 and 3, which show the mean accuracy or accuracy for each condition across the random seeds. Figure 2 shows these results for the Baseline, and Figure 3 shows them for the GogateCNN.

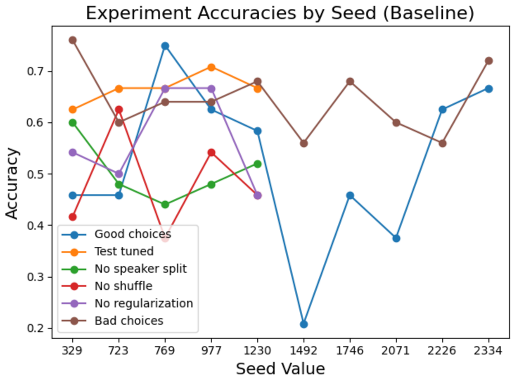

Figure 2: Average accuracy achieved for each seed and each condition, for the Baseline model.

Lastly, the results of the random seed analysis can be found in Table 4. The mean, min, and max accuracy across random seeds is shown for both models, and for 10 random seeds or 100 random seeds.

6 Analysis

6.1 Failed replication

The original result in Gogate et al. (2017) was far from being achieved with our setup, when proper methodology was used. Two opposing hypotheses for this are: 1) their result is legitimate, but

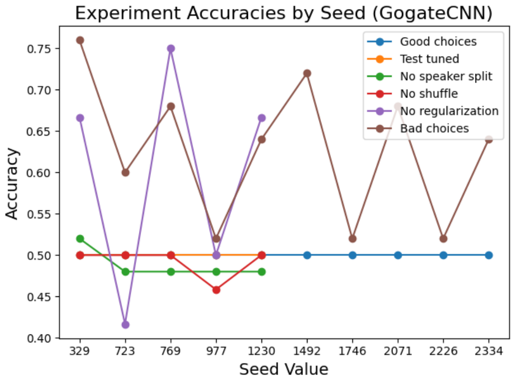

Figure 3: Average accuracy achieved for each seed and each condition, for the GogateCNN model.

Model	Num. seeds	Mean/max/min	Accuracy
Baseline	10	Mean	53.33%
Baseline	10	Min	33.33%
Baseline	10	Max	70.83%
Baseline	100	Mean	52.96%
Baseline	100	Min	29.17%
Baseline	100	Max	70.83%
GogateCNN	10	Mean	50.00%
GogateCNN	10	Min	50.00%
GogateCNN	10	Max	50.00%
GogateCNN	100	Mean	50.00%
GogateCNN	100	Min	50.00%
GogateCNN	100	Max	50.00%

Table 4: Validation accuracies across 10 data splitting random seeds, for untrained models. The high variance for the Baseline may be indicative of a lack of independence in the dataset. The low variance for the GogateCNN may be a result of the complexity of the model.

something about our approach was flawed; for instance, perhaps the Bayesian Optimization setup was flawed in a way that prevented it from learning performant hyperparameters; 2) their result is not legitimate, and one or more methodological flaws made it possible; for instance, if the random seed for data splitting was not explicitly fixed, it could potentially change across runs. The data above suggests this would lead to large variations in performance, and so would eventually leading to performant results without the authors necessarily knowing the true cause.

6.2 Ablations

The ablation studies had mixed results. Notably, standard deviation was high across random seeds. As a result, taking the max across random seeds consistently led to performance gains. Test tuning also led to considerable gains with the Baseline, whereas lack of regularization led to considerable gains in the GogateCNN. Also notable, the GogateCNN predicted exactly 50% for several runs, indicating as issue. Also, for both the Baseline and the GogateCNN, making only poor methodological choices consistently led to significant performance gains. Lastly, from the plots of accuracies by random seed, it seems there is correlation between performance levels by random seed, suggesting perhaps that some seeds are generally “better” than others.

6.3 Random Seed Analysis

Interestingly, the untrained Baseline model had very large variations in accuracy by random seed, a 37% difference between the max and min across only 10 random seeds. If we were to assume the untrained model is equivalent to random guessing, and that the validation samples are independent, then the probability that the model would achieve 70.83% accuracy or better is only 3%, which comes from applying the cumultative distribution of the binomial distribution. Across 10 samples, at best this happening once has a 30% chance, using a union bound. So, from this data alone, it is unclear whether the assumption that the model approximates independent random guessing is warranted. If its not warranted, and its not a coincidence that this somewhat unlikely outcome occurred, that would suggest that that the data samples are not independent. For instance, perhaps if the randomly initialized model predicts a label for one sample, it is likely to predict the same label for a sample in the same "cluster." The clusters would be something inherent to the data, and for instance perhaps the data is clustered by speaker. In any case, more examination is needed to determine whether this may be the case. Since similar, but unreported results were produced, its expected these results are more than coincidence. If that turns out to be the case, it could be be a methodological flaw for all papers using the Real-Life Trial dataset.

As for the GogateCNN, it predicted 50% accuracy every time, which could be due to the complexity of the model. Specifically, due to the numerous layers, it could be that the input data essentially gets washed out, leaving only noise, and the noise adds up to be tightly centered such that each label is predicted almost exactly 50% of the time. More investigation would be needed to verify this, and to test other hypotheses.

7 Conclusion

The unimodal textual classifier from Gogate et al. (2017) was replicated, but achieved accuracy fell short by 34% when proper methodology was employed. Additionally, it was shown that improper methodology can lead to an accuracy boost of up to 26%, and that random seeds can account for variation of up to 37% across only 10 random seeds.

Ultimately, the fundamental issue with Gogate et al. (2017) seems to be both improper methodology, and the smallness of the dataset - and these two factors are some cause for doubt over the impressive results reported generally in the field of machine learning deception detection.

Lastly, more work is needed on the subject of validating the performance of machine learning deception detection systems. Similar to the polygraph, understanding the true reliability of these systems may be crucial to impactful decisions, whether policy decisions or investment decisions.

References

Ann Cavoukian and Ronald J. Heslegrave. 1980. The admissibility of polygraph evidence in court: Some empirical findings. *Law and Human Behavior*, 4(1–2):117–131.

Alex Sebastião Constâncio, Denise Fukumi Tsunoda, Helena de Fátima Nunes Silva, Jocelaine Martins da Silveira, and Deborah Ribeiro Carvalho. 2023. Deception detection with machine learning: A systematic review and statistical analysis. *PLOS ONE*, 18(2):1–31.

Mandar Gogate, Ahsan Adeel, and Amir Hussain. 2017. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6.

Farhad Maleki, Katie Ovens, Rajiv Gupta, Caroline Reinhold, Alan Spatz, and Reza Forghani. 2023. Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *Radiology: Artificial Intelligence*, 5(1):e220028.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI ’15, page 59–66, New York, NY, USA. Association for Computing Machinery.

Leonard Saxe and Gershon Ben-Shakhar. 1999. Admissibility of polygraph tests: The application of scientific standards post- daubert . *Psychology, Public Policy, and Law*, 5(1):203–223.

Leonard Saxe, Denise Dougherty, and Theodore Cross. 1985. The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist*, 40(3):355–366.

Gaël Varoquaux and Veronika Cheplygina. 2022. Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1).