Lying Through One's Teeth: A Study on Verbal Leakage Cues

Min-Hsuan Yeh and Lun-Wei Ku
Institute of Information Science, Acadimia Sinica
Taipei, Taiwan
{samuelyeh, lwku}@iis.sinica.edu.tw

Abstract

Although many studies use the LIWC lexicon to show the existence of verbal leakage cues in lie detection datasets, none mention how verbal leakage cues are influenced by means of data collection, or the impact thereof on the performance of models. In this paper, we study verbal leakage cues to understand the effect of the data construction method on their significance, and examine the relationship between such cues and models' validity. The LIWC word-category dominance scores of seven lie detection datasets are used to show that audio statements and lie-based annotations indicate a greater number of strong verbal leakage cue categories. Moreover, we evaluate the validity of state-of-the-art lie detection models with cross- and in-dataset testing. Results show that in both types of testing, models trained on a dataset with more strong verbal leakage cue categories—as opposed to only a greater number of strong cues—yield superior results, suggesting that verbal leakage cues are a key factor for selecting lie detection datasets.

1 Introduction

One theory of lie detection is about cues to lying: *Why and when do liars and truth-tellers display different behavior?* Ekman and Friesen (1969) proposed two categories of cues: deception cues and leakage cues. Deception cues relate to the content of lies, such as an inconsistency in one's story; leakage cues appear because liars' emotions betray their true feeling, which can be further classified into non-verbal and verbal leakage cues. Zuckerman et al. (1981) reject the utility of focusing on liars' emotions but link such cues to cognitive load, supported by Vrij et al. (2008, 2016, 2017). DePaulo et al. (2003) analyzes 158 cues to deception, including non-verbal and verbal leakage cues, finding that verbal leakage cues are more reliable than others. Studies such as Adams (1996), Smith (2001), and Levitan (2019) show that verbal leakage cues can be found through psycholinguistic dictionaries such as the LIWC lexicon (Pennebaker et al., 1999), LDI (Bachenko et al., 2008; Enos, 2009), and Harbingers (Niculae et al., 2015).

Many NLP studies have recently collected lie detection datasets and detected lies using computational models (Hirschberg et al., 2005; Pérez-Rosas et al., 2014; Peskov et al., 2020); most of these ignore traditional lie detection methods and findings, and have no follow-up studies, making it difficult to know which datasets are suitable for model training. To use machine learning approaches together with lie detection research in psychology and linguistics, and to seek a way to evaluate and select proper datasets, this study focuses on analyzing verbal leakage cues within; leakage cues hereafter indicate verbal leakage cues. We study leakage cues in terms of the data collection method and model performance. Seven lie detection datasets are adopted for experiments. We analyze these datasets using word categories defined in LIWC2015 (Pennebaker et al., 2015). Through this study, we aim to answer three questions: (1) How do data collection methods affect strong leakage cues? (2) What is the role of the quantity and the category of strong leakage cues in lie detection task? (3) Do strong leakage cues contribute to model validity? We expect these answers to help in the construction and selection of appropriate datasets for lie detection tasks.

2 Leakage Cues and Datasets

To understand how leakage cues contribute to lie detection, we first measure the extent of leakage cues in lie detection datasets.

2.1 Datasets

We consider seven lie detection datasets:

- **Diplomacy (DM)** (Peskov et al., 2020): conversation logs collected from Diplomacy, an online text-based board game.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4504–4510
November 7–11, 2021. ©2021 Association for Computational Linguistics

Dataset	Size	Label	Statement	Context?
DM	13,132	Lies	T	Yes
MS	9,676	Liar	T	Yes
OD	7,168	Lies	T	No
LIAR	4,560	Lies	A, T	No
BOL	502	Lies	A	Yes
MU3D	320	Lies	A	Yes
RLT	121	Liar	A	Yes

Table 1: Lie detection datasets used here. T: Textual. A: Audio. Audio statements in datasets are transcribed.

Category	Sample words
Discrep	could, would, if, need, should, want
Number	first, half, once, one, two, million*
Certain	all, always, every, never, sure
Differ	but, than, or, exclude, opposite*, other
Affiliation	we, together, friend, family, met
Quant	more, percent, some, any, bunch

Table 2: LIWC dictionary samples.

- **Mafiascum (MS)** (de Ruiter an, 2018): logs from Mafia, an online text-based board game.
- **Open_domain (OD)** (Pérez-Rosas and Mihalcea, 2015): collected using Amazon Mechanical Turk (AMT) by asking turkers to provide seven lies and seven truths in text.
- **LIAR** (Wang, 2017): audio and textual truth and false statements made by politicians collected from PolitiFact.com[1].
- **Box_of_lies (BOL)** (Soldner et al., 2019): video clips collected from Box of Lies games from Tonight Show Starring Jimmy Fallon.
- **MU3D** (Lloyd et al., 2018): audio statements about social relationships collected in laboratory interviews.
- **Real-life_trial (RLT)** (Pérez-Rosas et al., 2015): true and false courtroom testimonies collected from InnocenceProject.org[2].

Dataset statistics are provided in Table 1.

2.2 Datasets and Dominant Word Categories

We start by measuring the representation of leakage cues in datasets using word-category dominance scores. The word categories $\mathcal{C}$ used here are defined in LIWC2015. LIWC is a psycholinguistic dictionary that groups words into 93 categories relevant to psychological processes, which has been used to detect leakage cues in multiple deception studies (Newman et al., 2003; Ott et al., 2011). Example LIWC word categories and their words are given in Table 2.

To calculate the dominance score of a word category $C_i \in \mathcal{C}$ in a dataset D, we first divide the samples in D into lie set L and truth set T. We calculate the lie and truth coverage rate of C_i as

$$u_i^L = \frac{1}{|L|} \sum_j^{|C_i|} v(L, i, j) \tag{1}$$

$$u_i^T = \frac{1}{|T|} \sum_j^{|C_i|} v(T, i, j), \tag{2}$$

where $v(\,\cdot\,, i, j)$ measures the occurrence count of word $w_{i,j} \in C_i$ within a given set. $|L|$ and $|T|$ represent the number of tokens in L and T, respectively. The dominance score of C_i is calculated by

$$r_i = u_i^L / u_i^T. \tag{3}$$

An $r_i \geq 1.2$ indicates a more deceptive category; an $r_i \leq 0.8$ indicates a more truthful category. In both cases, C_i is a strong word category (Mihalcea and Strapparava, 2009). Thus, we define a set of strong word categories

$$\mathcal{S} = \{C_i \mid r_i \geq 1.2 \vee r_i \leq 0.8\} \tag{4}$$

and the number of dominant words as

$$\rho = \sum_i^{|\mathcal{S}|} \sum_j^{|C_i|} v(D, i, j). \tag{5}$$

We refer to the number of dominant words as the number of strong leakage cues, and the number of strong word categories as the number of strong leakage cue categories.

Similar to Levitan (2019), LIWC word categories that cover less than 1% of truths or lies are first removed to minimize noise. LIWC word categories related to punctuation are also removed for normalization as this is not included in some transcriptions. The number of remained LIWC word categories is denoted as $|\mathcal{C}|$.

2.3 Analysis by Dataset

As shown in Table 2, studies collected lie detection data using various approaches. We are interested in how data collection methods affect leakage cues, specifically, how we can construct datasets to obtain more leakage cues for model learning. We list in Table 3 the max. and the min. dominance scores, their differences, the number of strong categories

[1] https://www.politifact.com/
[2] https://innocenceproject.org/

Dataset	Max	Min	Max-Min	$\lvert\mathcal{S}\rvert$ / $\lvert\mathcal{C}\rvert$	ρ
MS	1.05	0.96	0.09	0 / 44	0
DM	1.20	0.78	0.42	3 / 43	**19,180**
MU3D	1.38	0.72	0.66	6 / 50	3,952
OD	2.11	0.83	1.28	6 / 48	4,408
LIAR	1.31	0.68	0.63	10 / 50	11,647
RLT	2.36	**0.32**	2.04	18 / 52	2,338
BOL	**4.55**	0.36	**4.19**	**44 / 52**	2,823

Table 3: Dataset aspect analysis. Max, Min: the maximum and minimum dominance score. Max-Min: the differences between Max and Min. $|\mathcal{S}|$: the number of strong leakage cues categories. $|\mathcal{C}|$ the number of filtered LIWC word categories. ρ: the number of dominant words.

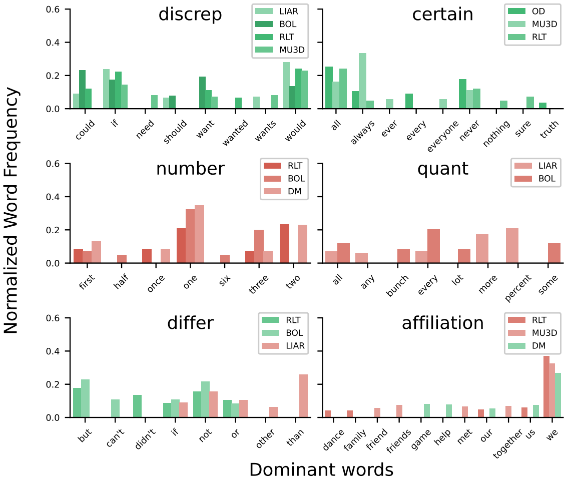


Figure 1: Normalized word frequency of the most common categories among seven datasets. For each dataset, we plot the normalized word frequency of the top five frequent words in each category. In this figure, distributions of top-5-word in 6 categories are shown. Green (Red) bars refer to deceptive (truthful) categories.

and total categories, and the number of dominant words. Their analyses are shown below.

Audio vs. Textual Sample Datasets with audio recordings (BOL and RLT) have many strong word categories: 44 for BOL and 18 for RLT, whereas data collected from textual statements (MS, DM, and OD) have few strong word categories. This may be due to the difference in complex cognitive load, as Zuckerman et al. (1981) shown in their work. Comparatively, deceiving using text incurs a smaller cognitive load, given that when typing, liars have more time to think and no need to control nonverbal behaviors. The only exception is audio-recorded MU3D, which asks interviewees to record four statements honestly and dishonestly about their social relationships. Since they can prepare the statements beforehand, and interviewers do not predict which statement is true, their cognitive load may not be as heavy as other audio datasets where lies are generated on the fly.

Lie- vs. Liar-based Annotation Results also show that datasets with liar-based annotation have few strong leakage cue categories. Comparing textual datasets, there is no strong leakage cue category in liar-based MS; comparing audio datasets, the number of strong leakage cue categories in liar-based RLT is considerably fewer than in lie-based BOL. Note that liars tend to wrap their lies with true information in an effort to be convincing (Peskov et al., 2020), i.e., lies are diluted by truth.

2.4 Analysis by Word Category

We find 53 strong word categories from 7 experimental datasets; 10 of these dominate on more than 3 datasets. To dig deeper into each category, for each dataset, we calculated the normalized word frequency for $w_{i,j}$—the proportion of the $w_{i,j}$ word frequency to the total words that belong to C_i:

$$\phi_{i,j} = v(D, i, j) / \sum_{k}^{|C_i|} v(D, i, k). \tag{6}$$

We plot the normalized word frequency of the most common strong leakage cues categories in Figure 1 and give some examples of lies and truths with those salient cues in Table 4.

In general, word categories capture words that are frequently used in lying and truth-telling. The upper results in Figure 1 show that liars in most datasets consistently use both *Discrep* and *Certain* words, suggesting that when lying, people tend to obscure facts with subjunctive mood, but also attempt to use definite words to increase credibility (See the first and second lies in Table 4). The middle results in the figure show that liars seldom use *Number* and *Quant* words, suggesting that liars are unwilling to include details (See the first and second truths in Table 4 add for camera ready). This supports the cognitive theory: describing details is a cognitively complex task.

In some cases, word categories dominate on different sides in different datasets. For example, *Differ* is a truthful category in LIAR but is a deceptive category in BOL and RLT. The lower-left results show that words in the *Differ* category used in LIAR are different than those used in RLT and BOL, suggesting that word usage affects which side

	Examples
Lies	*Says* he *never said* he *would* keep education funding the same. (LIAR) *No* sir I did *not*. I *absolutely* did *not*. *No* sir I was *not*. *No* sir. (RLT) *We're friends*, right? I *believe* that every message I've sent you has been truth. Are *we* still *friends*? (DM)
Truths	*One* is dusty. *One*'s got big hair. *One*'s got *hundreds* and *thousands* on it. (BOL) Weve created *more than* 850,000 jobs, *more than all* the *other* states combined. (LIAR) ... *we* would *share* a room *together* even though *we* had *our* own separate bedrooms... (RLT)

Table 4: Some examples of lies and truths. We select 3 lie and truth samples from the top 10 samples with many dominant words. Dominant words here are marked with *italic*.

the category is dominant on. Another interesting reason to cause a word category to dominate in both lies and truths is the nature of the scenario. *Affiliation* is a deceptive category in DM but a truthful category in RLT and MU3D. The lower-right results show that *Affiliation* words are used in these three datasets in a similar way: *we* is used frequently. However, data in DM are collected from a board game, and people in that game tend to deceive others when they are in alliances (See the third lie in Table 4). Therefore, *Affiliation* dominates on different sides in these three datasets. These results suggest that word categories provide insights to how humans lie in different scenarios.

Model	Inconsistent Label Rate						
	MS	DM	MU3D	OD	LIAR	RLT	BOL
BERT	0.86	**0.16**	1.0	**0.26**	**0.21**	**0.24**	0.58
CNN	0.82	0.52	1.0	0.58	0.53	0.62	0.42
UniGRU	**0.75**	0.43	1.0	0.61	0.48	1.0	**0.36**

Table 5: Inconsistent label rate for each model trained on different datasets (lower is better). Models with an inconsistent label rate of 1 would be excluded from further experiments.

3 Leakage Cues and Model Validity

To explore the effect of leakage cues on model validity, we conducted both cross- and in-dataset evaluations. We adopted three lie detection models for the experiments: UniGRU (an, 2014), CNN (Zhang and Wallace, 2017), and BERT (Devlin et al., 2019).

3.1 Experimental Details

For lie-based datasets, as each sample is an utterance labeled as lie or truth, we do not consider speakers of samples while splitting them into train/eval/test sets. For liar-based datasets, on the other hand, we concatenate all samples of one speaker to be one sample and assign it a speaker-level annotation in preprocessing before we split these samples.

For all three models, we use Adam with a lr of 3e-4 as the optimizer, and set the maximum number of input tokens to 256 as 95% of the sample's length is smaller than this number except the liar-based samples. For liar-based samples, some samples are too long to input into the NN model. As we found that the F1 score of those samples has no significant difference when configuring the maximum number of input tokens as 256 or 512, we keep the same setting and use 256 here. We apply batch sizes from 10 to 300 for different datasets, depending on their training set size. To deal with the unbalanced labels in some datasets, we apply weighted binary cross-entropy loss and use the ratio of labels as the weight.

3.2 Reliability

To discuss model validity, we first test the reliability and remove unreliable settings. We seek to evaluate model validity only on reliable models, that is, models with low label inconsistency rates. We define this rate as the percentage of samples that are sometimes predicted as true but sometimes as false by models trained with different random seeds. We trained 50 models with different random seeds as $M = \{m_1, ..., m_{50}\}$ and tested them on testing set D^{ts}, where $d \in D^{ts}$. The inconsistent label rate ϵ is measured as

$$\epsilon = |D^t \cap D^f| \; / \; |D^{ts}| \tag{1}$$

$$D^t = \{d | \exists \; m \in M, f(d; m) \geq 0.5\} \tag{2}$$

$$D^f = \{d | \exists \; m \in M, f(d; m) < 0.5\}. \tag{3}$$

A smaller ϵ indicates lower label inconsistency.

Table 5 shows that all three models trained on MU3D and UniGRU trained on RLT yielded an ϵ of 1, indicating models with highly inconsistent labels; we excluded these from further analysis.

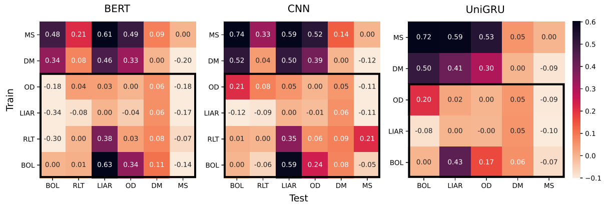
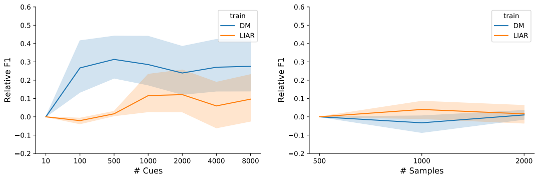


Figure 2: Experiment result of inter-dataset validity (lower is better). Datasets are ordered from BOL to MS by the increasing number of strong leakage cues categories (See Table 3).



(a) Fixed sample size

(b) Fixed number of cues

Figure 3: Experiment result of inner-dataset validity. The shade is the standard error, and the solid line is the mean of the results of the 7 datasets we tested in this experiment.

3.3 Validity

Inter-dataset Validity In this experiment, we examined how the number of strong leakage cue categories affects model validity when training on one dataset and testing on another.

Inspired by (Chen et al., 2020), inter dataset validity is measured as $\text{F1}_{drop} = \text{F1}_{in} - \text{F1}_{cross}$: a small F1_{drop} across datasets indicates good validity. Results in Figure 2 show that for each model, F1_{drop} in the black cluster is smaller than others, indicating that training on datasets with many strong leakage cue categories (BOL, RLT, LIAR, and OD) yields good or even better testing results on other datasets, i.e., good inter-dataset validity. Accordingly, to acquire a generalizable model, a lie detection dataset containing many strong leakage cue categories should be selected.

Inner-dataset Validity In this experiment, to understand how model performance changes, we controlled the training set (1) by varying the number of strong leakage cues while fixing the dataset size, and (2) by varying the dataset size while fixing the number of strong leakage cues. Models used here are UniGRU, and datasets are DM and LIAR, which include samples with many strong leakage cues.

We first set the dataset size to 1,000 samples, and use seven different numbers of strong leakage cues. The result is shown in Figure 3a. Models trained on datasets with many strong leakage cues (DM, blue) yield a significantly high F1 score. In particular, models trained on DM improve the F1 score by more than 40% when the number of strong leakage cues in the training set increases from 10 to 100. This result also shows that the F1 score increases with the number of strong leakage cues in training, and that this increase ceases after the number of strong leakage cues exceeds 1,000.

To evaluate the impact of dataset size, we fixed the number of strong leakage cues to 2,000, and used three different dataset sizes. As shown in Figure 3b, models trained on datasets with many samples also achieve a high F1 score on some settings, whereas this improvement is less compared to when the number of strong leakage cues is increased. Moreover, in some settings, performance fails to improve when the dataset size is increased. These two experiments suggest that the number of strong leakage cues in datasets is more critical for model validity than the dataset size. Therefore, we argue that a good lie detection dataset should contain many strong leakage cues.

4 Conclusion

In this paper, we study the convolutions among leakage cues, datasets, and models. Various conditions are analyzed, with results that show that leakage cues help increase model validity, and that they can be found the most in datasets containing audio statements and lie-based annotations. These findings and the testing methods are good references for selecting appropriate data and models when building lie detection applications. Under the condition that no benchmark has been recognized yet, we expect this research to serve as a guide for researchers new to this problem, saving them unnecessary effort and helping them to quickly get up to speed.

5 Ethical Considerations

We analyze the relationship between verbal leakage cues and existing lie detection datasets and models, providing a proper way to select and collect lie detection data. We found that a good lie detection dataset should contain many strong leakage cue categories, which can be achieved with audio statements and lie-based annotation, not related to race, sex, or other factors which may cause ethical issues. We believe that this study can help improve the quality of lie detection datasets and models, and protect people from deceived.

Acknowledgements

This research is partially supported by Ministry of Science and Technology, Taiwan under the project contract 108-2923-E-001-001-MY2 and 110-2221-E-001-001-.

References

S. H. Adams. 1996. Statement analysis: What do suspects' words really reveal?

Junyoung Chung an. 2014. Empirical evaluation of gated recurrent neural networks on sequenc. *ArXiv preprint*, abs/1412.3555.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 41–48, Manchester, UK. Coling 2008 Organizing Committee.

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691, Online. Association for Computational Linguistics.

Bob de Ruiter an. 2018. The Mafiascum Dataset: A Large Text Corpus for Deception Detection. *ArXiv preprint*, abs/1811.07851.

Bella DePaulo, James J Lindsay, Brian Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129:74–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman and Wallace V. Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106. PMID: 27785970.

Frank Enos. 2009. *Detecting Deception in Speech*. Ph.D. thesis.

Julia Hirschberg, S. Benus, Jason M. Brenier, Frank Enos, S. Friedman, S. Gilman, Cynthia Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. In *INTERSPEECH*.

S.I. Levitan. 2019. *Deception in Spoken Dialogue: Classification and Individual Differences*. Columbia University.

E. Paige Lloyd, Jason Deska, Kurt Hugenberg, Allen McConnell, Brandon Humphrey, and Jonathan Kunstman. 2018. Miami University deception detection database. *Behavior Research Methods*, 51.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore. Association for Computational Linguistics.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675. PMID: 15272998.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1650–1659, Beijing, China. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015.

James Pennebaker, Martha Francis, and Roger Booth. 1999. Linguistic Inquiry and Word Count (LIWC).

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. Association for Computing Machinery.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, Lisbon, Portugal. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Alexis Narvaez, and Mihai Burzo. 2014. A multimodal dataset for deception detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3118–3122, Reykjavik, Iceland. European Language Resources Association (ELRA).

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Nicky Smith. 2001. Reading between the lines: An evaluation of the Scientific Content Analysis technique (SCAN). *Police Research Series Paper*, 135.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics.

Aldert Vrij, Ronald Fisher, Samantha Mann, and Sharon Leal. 2008. A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, 5(1-2):39–43.

Aldert Vrij, Ronald P. Fisher, and Hartmut Blank. 2017. A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.

Aldert Vrij, Ronald P. Fisher, Hartmut Blank, Sharon Leal, and Samantha Mann. 2016. *A cognitive approach to elicit verbal and nonverbal cues to deceit*, page 284–302. Cambridge University Press.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Ye Zhang and Byron Wallace. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and Nonverbal Communication of Deception11. volume 14 of *Advances in Experimental Social Psychology*, pages 1 – 59. Academic Press.