Box of Lies®: Multimodal Deception Detection in Dialogues

Felix Soldner
Department of Security
and Crime Science
University College London
felix.soldner@ucl.ac.uk

Verónica Pérez-Rosas and **Rada Mihalcea**
Department of Computer Science
and Engineering
University of Michigan, Ann Arbor
{vrncapr,mihalcea}@umich.edu

Abstract

Deception often takes place during everyday conversations, yet conversational dialogues remain largely unexplored by current work on automatic deception detection. In this paper, we address the task of detecting multimodal deceptive cues during conversational dialogues. We introduce a multimodal dataset containing deceptive conversations between participants playing The Tonight Show Starring Jimmy Fallon® Box of Lies game, in which they try to guess whether an object description provided by their opponent is deceptive or not. We conduct annotations of multimodal communication behaviors, including facial and linguistic behaviors, and derive several learning features based on these annotations. Initial classification experiments show promising results, performing well above both a random and a human baseline, and reaching up to 69% accuracy in distinguishing deceptive and truthful behaviors.

1 Introduction

Deception occurs often during dialogues, but until now this setting has received little attention from the research community (Tsunomori et al., 2015). In this paper, we explore verbal, non-verbal, and conversational dialog cues between contestants playing the Box of Lies game in The Tonight Show Starring Jimmy Fallon® tv show. In the game, participants try to guess whether an object description provided by their opponent is deceptive or not. The game scenario provides a rich environment where we can explore several aspects of deceptive behavior occurring during conversations. First, it allows us to study conversational deception in the presence of multiple modalities such as verbal and non-verbal behaviors. Second, it provides observable assessments of participant's honesty, which is usually an important challenge during deception research. Third, since participants experience the pressure to win the game in front of a big audience, it presumably presents an environment with high stakes.[1]

Recent work on multimodal deception detection has already shown the importance of verbal and non-verbal behaviors during the automatic identification of deceit (Abouelenien et al., 2017a). Following this line of work, our main contribution consists of investigating whether such modalities can also be leveraged to predict deception in a conversational dialog as well as exploring whether the dialogue setting adds meaningful information to the other modalities to potentially increase the classification performance. Based on earlier work (Perez-Rosas et al., 2014; Mihalcea et al., 2013), we hypothesize that **(1)** including dialogue features in addition to other multimodal features (language and facial expressions) while training a classifier increases the prediction performance; **(2)** automatic classification of truthful and deceptive behavior is better than random guessing (50% for equal class sizes); and **(3)** automatic classification of truthful and deceptive responses is more accurate than human judgments (based on performance of participants in the dataset).

To address these hypotheses, we first generate a dataset containing verbal and non-verbal annotations of deceptive and truthful interactions between the game participants. Next, we derive linguistic, visual, and dialog cues based on our annotations for the verbal and non-verbal components of the dataset. The features are then used to conduct several learning experiments under different scenarios that attempt to distinguish between deceptive and truthful utterances either by focusing on the statements generated by one participant at a time (the game's host or the guest), or by address-

[1] To our knowledge, these conversations are not scripted.

ing them all together.

Our initial experiments show that language, as well as behavioral and dialog features, carry meaningful information. Moreover, the automatic classification of deception can be performed with an accuracy that is better than random guessing and ourperforms human judgments.

2 Related Work

To tackle the problem of reliably detecting deception, researchers have applied various forms of automated deception detection methods that rely on machine learning approaches, which are able to incorporate a variety of behavioral cues from text, audiovisual or physiological data sources (Ott et al., 2011; Fornaciari and Poesio, 2013; Mihalcea and Strapparava, 2009; Abouelenien et al., 2016).

Many studies focused on text-based classification, detecting false online reviews (Ott et al., 2013) or deceptive transcribed statements from court hearings (Fornaciari and Poesio, 2013). Other studies utilized visual cues such as facial expressions or other body movements to detect deception (Meservy et al., 2005). These methods already show success in identifying deceptive behavior using individual modalities. In addition, recent approaches, which combine multiple modalities are able to further boost classification performances (Abouelenien et al., 2017b; Perez-Rosas et al., 2015).

However, multimodal approaches have not utilized the dialogue dimension in combination with other modalities as of yet. The dialogue dimension captures the interaction between two individuals and how they react to each other. One previous study investigated such an interaction, in which the researchers examined question types and their behavioral effect on participants (Tsunomori et al., 2015). Findings of the study showed that specific questions led to more salient deceptive behavior patterns in participants. This increase in feature salience resulted in better deception detection performances.

The interaction between two individuals in deceptive conversations was also investigated by Hancock et al. (2004), who examined deception at the linguistic level. Participants, who were unaware of receiving a deceptive message, produced more words, sense terms and asked more questions as compared to when they received a truthful message. In a similar setting, in which two participants engaged in a question-response task, Levitan et al. (2018) examined linguistic, gender and native language differences. They found significant variations in these features for truthful and deceptive responses. The experimenters utilized these variations in an automated classification task and reached up to 72% accuracy.

These studies show that a focus on the linguistic level and the interaction between individuals can have a beneficial effect on detecting deceit. Other studies examined non-verbal behavior. In an experiment, Sen et al. (2018) video-recorded conversations between participants in an interrogation game and examined participant's facial expressions. The results showed that interrogators exhibited different facial expressions when they were lied to as opposed to when they were told the truth. In a different approach, Yu et al. (2015) observed head movements and facial expressions between two individuals. The authors established normalized non-verbal patterns, which enabled them to capture interactional synchrony. This allowed them to successfully discriminate between truths and lies in the experiment.

Overall, this previous research demonstrates that capturing verbal or non-verbal interactions can convey meaningful information about deceit, which can be leveraged for multimodal deception detection.

3 Dataset of Deceptive Conversations

To explore the role played by conversation dynamics in deceptive behaviors, we collected conversations where participants acted deceptively. Specifically, we opted for identifying public sources where the veracity or falsehood of conversation participants is known.

In the game show Box of Lies, which is part of the late-night talk show The Tonight Show Starring Jimmy Fallon, these labels are known. The host (Jimmy Fallon) and his guest play the game Box of Lies, where participants take turns to play the game. During the game, when is their turn, the participants pick a box (from among nine available boxes) that contains an object they have to describe to their opponent. The object is hidden from the opponent through a separation wall between the two contestants. Participants sit opposite to each other and see their upper body and face through a cut hole in the separation wall. The opponent must guess if the provided description is

truthful or not. The participant with the best of three guesses wins the game.

This setup allows us to observe verbal and non-verbal behavior exhibited by the participants during the dialogue interaction. In order to better capture multimodal behavioral cues of deception throughout the conversation, we decided to conduct annotations at utterance-level. We thus built a rich multimodal dataset containing verbal and non-verbal annotations for 1049 utterances, which is used in the experiments reported in this paper. The data collection and annotation process are described below.

3.1 Data Collection

We search for publicly available Box of Lies videos on the YouTube platform.[2] We collected 25 videos that are currently available in the show video-feed. The full set consists of 2 hours and 24 minutes of video. The average length of a video is six minutes and contains around three rounds of the game (this varies depending on the score and on whether additional time was available for extra rounds). Each video features a different guest and Jimmy Fallon, resulting in 26 unique participants, with 6 of them being males and 20 females.

3.2 Annotation of Multimodal Communication Behaviors

To capture the non-verbal behavior of the participants, each video is initially segmented based on the conversation turn-taking and annotated with the help of the ELAN software (Wittenburg et al., 2006). ELAN provides a multimodal annotation platform on which audiovisual recordings are annotated in a multi-level tier structure. In our case, we defined the following structure to annotate both types of behavior: host verbal, host non-verbal, guest verbal, and guest non-verbal.

3.2.1 Non-verbal Behaviors

To annotate facial and communication behaviors, we use MUMIN, a multimodal coding scheme that is used to study gestures and facial displays in interpersonal communication with a focus on the role played by multimodal expressions for feedback, turn management, and sequencing (Allwood et al., 2005). Given the nature of the video-conversations being depicted in our dataset, which show the face and upper bodies of the participants and their interaction, we focus our annotations on facial and conversational behavior. These choices are motivated by previous research showing that different expressions for truthful and deceptive behaviors are present (DePaulo et al., 2003) in the eyes and mouth regions, as well as studies on the role of conversational involvement in deceptive interactions (Burgoon et al., 1999).

Facial behaviors. We annotate the categories for visual cues and behaviors of eyebrows, eyes, gaze, mouth-openness, mouth-lips, head, and the general face. Each of the categories takes on one of several mutually exclusive behavior values. Table 1 shows the frequencies of all facial expressions included in this set. In the table, we observe a slightly unequal representation of behavioral categories (e.g., head movements are observed more often than other facial expressions). This is mainly attributed to camera angle changes during the videos causing participant's faces to be only partly or not visible, thus restricting the behavioral coding. The annotated values reflect the most dominant observed behavior in that time segment of the video.

Two annotators coded the videos, and after the first three videos, the inter-annotator agreement was measured by calculating the Kappa score, to ensure accurate coding. If the agreement was below Kappa (weighted) = 0.45 in any category, this category was discussed to identify and reconcile differences in the coding strategy. The annotators re-coded the videos individually and compared them again. This process was repeated until the desired agreement was reached (above .40 for each category). In most cases, we repeated the process only twice, except for the "feedback receiving" and "feedback eliciting" categories which were discussed three times. Table 2 shows the final Kappa score for each category.

3.2.2 Speaker's Veracity

In the full video set, participants play 68 rounds (29 truthful and 39 deceptive). Occasionally, deceptive rounds also contain truthful statements, in which contestants describe parts of the object truthfully, but other parts deceptively, turning the overall description into a lie. For example, a contestant might say: "I have before me, a green lobster on a plate." In truth, the object is a red lob-

[2]The videos are originally produced by NBC and retrieved from Youtube. We consider that using YouTube videos for research purposes falls under the "fair use" clause, which is stated on: https://www.youtube.com/intl/en-GB/yt/about/copyright/fair-use/

Label	Count	Label	Count	Label	Count
General face		**Head**		**Mouth-Openness**	
Smile	411	Neutral/still	320	Open mouth	763
Neutral	342	Waggle	292	Closed mouth	212
Other	100	Side-turn	242	Other	3
Laughter	83	Single Nod (Down)	165	**Gaze**	
Scowl	42	Move Forward	153	Towards interlocutor	674
Eyebrows		Repeated Nods (Down)	122	Towards object	148
Neutral/Normal	531	Move Backward	117	Down	37
Raising	320	Single Tilt (Sideways)	115	Towards audience	36
Frowning	76	Single Jerk (Backwards Up)	78	Sideways	35
Other	39	Shake (repeated)	75	Other	34
Mouth-Lips		Repeated Tilts (Sideways)	18	**Eyes**	
Retracted	279	Other	15	Neutral/Open	465
Neutral	267	Single Slow Backwards Up	10	Closing-repeated	203
Corners up	261	Repeated Jerks (Backwards Up)	7	Closing-both	166
Other	102			Other	121
Protruded	46			Exaggerated Opening	8
Corners down	21			Closing-one	4

Table 1: Frequency counts for participants' face, head and mouth annotations.

Figure 1: Sample screenshots of truthful and deceptive behavior from the original videoclips; left-top (truthful) : Eyebrows-raising, Eyes-open; left-bottom (truthful): Eyebrows-neutral, Eyes-open; right-top (deceptive): Eyebrows-frowning, Eyes-closing (both); right-bottom (deceptive): Eyebrows-raising, Eyes-closing (both).

ster on a plate. The description contains truthful and deceptive aspects, but it is considered to be a deceptive round since the main purpose of the statement is to deceive. This fine-grained distinction is captured during the annotation of behaviors, described below, which allows us to obtain more precise veracity labels of the behavior. In our example, the behavior associated with the description "green" is labeled as deceptive, whereas all the other behaviors are labeled as being truthful.

To enable this annotation, we further process our initial turn-by-turn segmentation to obtain spoken segments by either of the participants. We then code the veracity (i.e., truthful or deceptive) for each verbal statement of the participants. During the veracity coding, we assume that the behavior is always deceptive unless the verbal description indicates otherwise (i.e., accurate description of the object), as the general goal of each participant is to deceive their opponent. The final distribution of these annotations is 862 utterances labeled as deceptive, and 187 as truthful. Figure 1 shows examples of truthful and deceptive behaviors in the dataset.

3.3 Transcriptions

In order to include linguistic features in our analyses, we first transcribe the participants' conversations. To obtain transcriptions, we first extract the

Category	Kappa	
	Host	Guest
General Face	0.75	0.70
Eyebrows	0.51	0.70
Eyes	0.56	0.92
Gaze	0.45	0.74
Mouth-Openness	0.64	0.47
Mouth-Lips	0.79	0.53
Head	0.60	0.55
Feedback receiving	0.47	0.72
Feedback eliciting	0.73	0.46
Average	0.61	0.64

Table 2: Inter-annotator agreement

	Truthful	Deceptive	Total
Host	749	4211	4960
Guests	748	2496	3244
Total	1497	6707	8204

Table 3: Distribution of words for all transcriptions

audio of the corresponding video clip and slice it based on the verbal annotation time-stamps. For this task, we use Pympi (Lubbers and Torreira, 2013) and Ffmpy (Developers, 2016). We transcribe the resulting audio clips using Amazon Mechanical Turk (AMT), a crowd-sourcing platform. We notice that some of the clips include brief interruptions among speakers, thus we ask the AMT workers to transcribe only the speech of the main speaker in the audio clip. After we collect all transcriptions, we proofread them to avoid mistakes such as double transcriptions and remove additional characters or descriptions (e.g. "person 1", clapping, [pause]). The final distribution of all the words from the transcriptions is shown in Table 3. Example utterances of truthful and deceptive statements are displayed in Table 4.

4 Methodology

Gathering data from different modalities creates the need to combine them into a coherent feature set, which can be utilized by machine learning classifiers. The following subsections describe how we generate features based on our annotations for the verbal and non-verbal behavior components of the dataset. These features are then used to train and test the classifiers in our experiments.

4.1 Linguistic Features

We derive various linguistic features from the transcriptions of the participants' speech, which include: unigrams, psycholinguistic features, part of speech features, and word embedding features.

Unigrams. These features are created with bag-of-words representations of all transcriptions from the guests and the host. The unigrams are represented using their frequencies.

Psycholinguistic Features. These features are created with the help of the Linguistic Inquiry and Word count Lexicon (Version 2015) (Pennebaker et al., 2007). They represent 80 different classes of words, which can be attributed to different psychological dimensions. The features display the frequencies of occurrences of classes, derived from the occurrences of words, attributed to each class. The lexicon has been successfully used in previous work for automatic deception detection (Ott et al., 2013; Mihalcea et al., 2013).

Part of Speech tags (PoS). These features are created by obtaining PoS-tagging of transcripts. They capture the grammatical and syntactical structure of the utterances of the transcriptions (e.g., noun, verb, adjective). Features display the distribution of these categories in percentage for each utterance.

Word Embeddings. These features are obtained using Word2Vec by creating vector representations of the words in the transcriptions. By training word representations based on other words occurring in the same context, these features capture similarities of words next to each other and in context. Together, all words are represented in a vector space in which similar words lay closer to each other as compared to dissimilar words.

4.2 Non-verbal Behavioral Features

These features are generated from the non-verbal behaviors described in Section 3.2 and represented as percentages. Specifically, the different behavioral values for a category (e.g., Head) in a verbal utterance are counted and represented as percentages. For example, a verbal utterance might last for one minute and during that time head movements might take several different values, such as *side-turn* (20 sec.), *shake* (30 sec.), and *single nod* (10 sec.). These times are transformed into percentages and the category *head* then consist of 33.33% *side-turn*, 50% *shake*, and 16.67% *single nod* during the one-minute utterance. In this manner, each facial area designates its percentage representation of behavioral values, which add up to 100%. In case a behavior cannot be fully attributed to one of the possible actions through the verbal statement, left-over percentages are assigned to

Participant	**Truthful**	**Deceptive**
Host	"In a, no. In a costume and also inside the box, a bunch of Hershey kisses."	"Ever heard of a boy band called Backstreet Boys?"
Guest	"Ok, it is, um, a rubiks cube inside jello."	"Okey. Its a toaster oven. Ohhh ha ha ha ha"

Table 4: Examples of utterances from the transcriptions

none, representing the lack of occurrence of a behavioral action in its category. This transformation is performed for all seven different facial areas we have annotated, including General Facial Expressions, Eyebrows, Eyes, Gaze, Mouth-Openness, Mouth-Lips, and Head.

Our non-verbal behavior feature set thus consists of all the facial expressions or head movements expressed as the percentage of times they occur during a speaker's utterance. Possible attributes for each of the seven categories can be found in Table 1.

4.3 Dialogue Features

We derive dialogue-based features by exploring verbal and non-verbal aspects of the interaction between participants that are related to deceptive behavior. The features attempt to capture deception cues the speaker's exhibited during their conversation prior to the current utterance. These features are obtained as follows:

Deception Changes. These features include the count of truthful and deceptive utterances up to the current utterance. We also aggregate the counts of deceptive and truthful utterances to represent the participation of the speaker during the conversation.

Non-verbal Changes. These features capture how facial displays differ between consequent utterances. We calculate these features by subtracting the numeric vectors representing the non-verbal behavior during the current utterance from the previous utterance.

4.4 Data Alignment

In order to attribute the corresponding non-verbal behaviors to verbal utterances for later classification tasks, each behavior receives a veracity label (truthful or deceptive) individually. The veracity label that overlaps with a behavior for more than 50% of its span, it is associated with that behavior. The overlap is determined by comparing the time stamp of the behavior and the veracity annotation, which are obtained from the ELAN files. Table 5 displays the distribution of these feature sets.

Participant	Lies	Truths	All
Guests	394	101	495
Host	468	85	553
All	862	187	1048

Table 5: Class distribution for host and guest features

5 Human Performance

In order to evaluate the automated methods and compare them to human performance, we establish a human baseline, representing how well humans guess deceptive and truthful behavior correctly. Since the game show Box of Lies is already set up in a way that participants have to guess if their opponent is lying or telling the truth, their performance serves as a baseline.

Thus, we use their assessments to obtain a confusion matrix showing their correct and incorrect guesses. We calculate their performance in terms of accuracy, which reflects the proportion of correctly categorized descriptions of all object descriptions; precision, which reflects the proportion of correctly identified descriptions in one classified category; recall, which reflects the proportion of correctly identified descriptions out of all the object descriptions truly belonging to that category; and f1-score, which reflects the weighted average of precision and recall in that category.

Human performance is shown in Table 6. Since participants tell 39 deceptive and 29 truthful descriptions in total, the distribution is slightly uneven, resulting in a baseline of 0.57 in detecting a lie. Considering this, participants and the overall accuracy is almost equal to the accuracy of random guessing. This supports earlier findings that humans are almost only as good as chance (Bond and DePaulo, 2006).

Results for each class (detecting truthful or deceptive descriptions) show that participants are better at detecting truthful descriptions. This could be based on the truth bias, which describes the phenomenon according to which people generally tend to believe others (Levine et al., 1999).

		Lie			Truth		
	Acc.	P	R	F	P	R	F
Guests	55%	0.67	0.26	0.38	0.51	0.86	0.64
Host	58%	0.17	0.17	0.17	0.72	0.72	0.72
All	56%	0.47	0.24	0.32	0.58	0.79	0.67

Table 6: Human performances in the game "Box of Lies"; Acc. = Accuracy; P = Precision; R = Recall; F = F1; Baseline (detecting a lie) = 0.57

6 Experiments

During our experiments, we use a Random Forest classifier. We perform all the classification experiments with the python package Scikit-learn (Pedregosa et al., 2011) using the standard settings for the model parameters. All classifiers are evaluated using five-fold cross-validation. During our experiments we focus on three scenarios:

(1) How well can we distinguish between truthful and deceptive utterances in the dataset? In this scenario, we explore whether the different features we propose can capture differences between truthful and deceptive behavior, regardless of the speaker. Note that in this scenario, a significant fraction of the data comes from the same speaker (host).

(2) How well can we distinguish between truthful and deceptive behaviors elicited by the guests? In this experiment, we consider the subset of deceptive and truthful utterances produced by the guests in our dataset. Again, we test our different feature sets in the prediction of truthful and deceptive behavior, but this time we focus on learning deceptive patterns from several individuals, which might exhibit different verbal and non-verbal behaviors.

(3) How well can we distinguish between truthful and deceptive behaviors exhibited by the host? In this scenario, we explore whether the availability of data by the same individual can help to improve the detection of deceptive behavior. In other words, this experiment builds personalized deception models for the host using the different sets of features representing verbal and non-verbal behavior.

For each scenario, we test classifiers with features derived from the different verbal and non-verbal modalities as well as features that represent the interaction between participants (described in Sections 4.1, 4.2 and 4.3). We test the predictive power of each feature set individually and we also build joint models that combine all feature sets. The classifiers performance is evaluated in terms of accuracy, precision, recall, and f1-score.

An important challenge during the experiments is that the nature of our dataset leads to a high unbalance between the truthful and deceptive classes, as shown in Table 5. During our experiments, the imbalance of the data is tackled by applying down-sampling to the deceptive class (Oliphant, 2006). This ensures an equal distribution of each label and results in a baseline of 0.50 in all scenarios. The results for verbal, non-verbal, dialog features, and their combination for each scenario are shown in Table 7.

Overall, our experiments show the benefit of combining multiple sources of information on this task, with accuracies well above the baseline and a noticeable accuracy improvement when using all feature sets.

7 Discussion

The different classification performances show that adding information from several modalities helps to increase the accuracy of the detection system. Not surprisingly, the linguistic modality shows the best performance among single modalities (Scenarios 1 and 2). More interestingly, the non-verbal modality is the second best indicator of deception, despite a significant amount of facial occlusions present in our dataset.[3] Furthermore, this finding is in line with other work on multimodal deception detection also showing that gestures are a reliable indicator of deception (Perez-Rosas et al., 2015).

In addition, we generate learning curves for each modality in scenario 1 (figure 2). The curves show that when training with 50 - 60% of the data, the classifier starts to improve upon the (guessing) baseline. The ascending trend does not seem to level off, even with the entire dataset, indicating that the classifier might benefit from more data.

Our experiment on exploring the classification of deceptive behaviors from the host (scenario 3) also lead to interesting insights. First, the linguistic modality is the weakest since it obtained the lowest performance in both classes. As the difference in f-score values shows, the host does not appear to use significantly different language while telling lies or truths, at least not at the lexical and semantic level, as captured by our linguistic features. Second, his non-verbal behavior does

[3]Facial occlusions are mainly attributed to changes in camera angles occurring during the videos

			Lie			Truth		
Scenario	Features	Acc.	P	R	F	P	R	F
(1) General truths and lies	Linguistic	62%	0.61	0.67	0.64	0.63	0.57	0.6
	Dialog	54%	0.54	0.56	0.55	0.54	0.52	0.53
	Non-verbal	61%	0.64	0.54	0.58	0.6	0.69	0.64
	All Features	65%	0.64	0.67	0.66	0.66	0.63	0.65
(2) Lies and truths by guests	Linguistic	66%	0.64	0.73	0.68	0.69	0.58	0.63
	Dialog	57%	0.58	0.52	0.55	0.56	0.61	0.59
	Non-verbal	61%	0.60	0.62	0.61	0.61	0.58	0.60
	All Features	69%	0.66	0.76	0.71	0.72	0.61	0.69
(3) Lies and truths by host	Linguistic	55%	0.55	0.60	0.57	0.56	0.51	0.53
	Dialog	58%	0.58	0.61	0.59	0.59	0.55	0.57
	Non-verbal	57%	0.56	0.62	0.59	0.58	0.52	0.55
	All Features	65%	0.67	0.61	0.64	0.64	0.69	0.67

Table 7: Automated performances for the three analyzed scenarios.The baseline for all scenarios (detecting a lie) is 50%; Acc. = Accuracy; P = Precision; R = Recall; F = F1-score

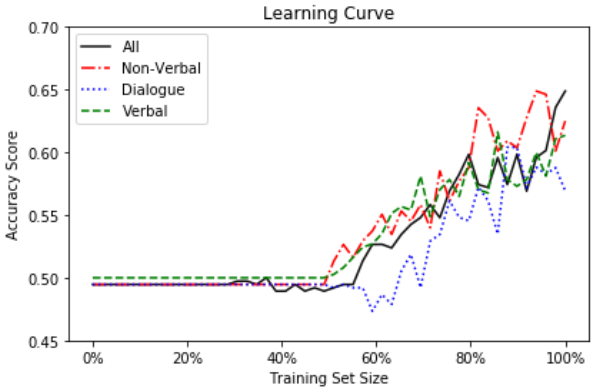

Figure 2: Learning curves (averaged 5-fold cross validation) for each modality using random forest (general truth/lie scenario)

seem to be different while telling lies and truths as we observe noticeable improvement in the models build with non-verbal cues. Third, the performance of the dialog features suggests that having evidence of previous behavior (as captured by deception changes and non-verbal behavior changes) can be useful when modeling the deceptive behavior of a single individual, further suggesting that the non-verbal component of a lie seems to be more individually shaped for each person as opposed to the linguistic component. However, the differences in performance between scenarios 1 and 2 suggest that the current features might not be enough to capture deceptive behavior by a single individual since the developed classifiers still find this task challenging.

The preliminary analyses of this new multimodal dataset show promising results by successfully classifying truthful and deceptive behaviors. Currently, the dataset provides feature sets drawn from three modalities (verbal and non-verbal, as well as dialogue) but can be further analyzed to extract additional features from other modalities such as speech. Specifically, the dialogue has the potential to add many more layers of information by systematically analyzing verbal, non-verbal, and speech patterns between the participants. These patterns can lead to detectable differences between actions and reactions within a dialogue (Tsunomori et al., 2015; Levitan et al., 2016). We consider analyzing such patterns as a future research venue to expand the analyses on our dataset.

A challenge while using this data is the current imbalance of truthful and deceptive feature sets, which can have a detrimental effect on classification performance. However, there are several other possible ways to address this issue other than down-sampling as we did during our experiments. For instance, other computational methods could be explored, such as one-class classification tasks. Such models train on a dataset from the same distribution and classify new data as being similar or different to that distribution. This way, anomalies (i.e., behavior with a feature configuration different from the training set) are detectable. Since truthful behavior is underrepresented in our dataset, the deceptive features could serve as the training set and the goal is to detect truthful behavioral patterns. Expanding on other computational tasks also tackles future applicable problems of dealing with uneven datasets, as they are often present when working with real-life datasets. The issue of an underrepresented class is prevalent in deception detection research.

Finally, the dataset could be expanded with more behavioral data, mainly by augmenting the number of truthful behaviors. Since all the contes-

tants in our dataset are celebrities, it is likely that other videos portraying them are available.

8 Conclusion

In this paper, we showed how we can successfully build a multimodal dialog dataset for deception detection, and presented exploratory deception classification tasks. We showed how we can integrate multiple modalities and build feature sets useful for automated processing. We were able to achieve a classification performance that is better than random guessing and exceeds human performance. Furthermore, additional modalities systematically showed an improvement in classification performance. The best performance of 69% was obtained by combining multiple verbal, nonverbal, and dialogue feature sets, which represents a significant improvement over the human performance of a maximum of 58% accuracy.

The dataset introduced in this paper represents a first attempt to integrate the dialogue dimension with multiple other modalities in deception detection research. It has the potential of triggering novel research on multimodal deception data, specifically for speech and the dialogue dimension, which should be explored in the future.

All the data annotations described in this paper are available upon request.

Acknowledgments

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science, the National Science Foundation, or the John Templeton Foundation.

References

Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2016. Analyzing Thermal and Visual Clues of Deception for a Non-Contact Deception Detection Approach. pages 1–4. ACM Press.

Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017a. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055.

Mohamed Abouelenien, Veronica Perez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017b. Detecting Deceptive Behavior via Integration of Discriminative Features From Multiple Modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055.

Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2005. The MUMIN multimodal coding scheme. *NorFA yearbook*, 2005:129–157.

Charles F. Bond and Bella M. DePaulo. 2006. Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Judee K. Burgoon, David B. Buller, Cindy H. White, Walid Afifi, and Aileen L. S. Buslig. 1999. The role of conversational involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin*, 25(6):669–686.

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118.

FFmpeg Developers. 2016. Ffmpeg tool (Version be1d324). http://ffmpeg.org/.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3):303–340.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael T Woodworth. 2004. Lies in Conversation: An Examination of Deception Using Automated Linguistic Analysis. page 6.

Timothy R. Levine, Hee Sun Park, and Steven A. McCornack. 1999. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, 66(2):125–144.

Sarah Ita Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 40–44.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic Cues to Deception and Perceived Deception in Interview Dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, Louisiana. Association for Computational Linguistics.

Mart Lubbers and Francisco Torreira. 2013. Pympiling: A Python module for processing ELANs EAF and Praats TextGrid annotation files.

T.O. Meservy, M.L. Jensen, J. Kruse, D.P. Twitchell, G. Tsechpenakis, J.K. Burgoon, D.N. Metaxas, and J.F. Nunamaker. 2005. Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior. *IEEE Intelligent Systems*, 20(5):36–43.

Rada Mihalcea, Verónica Pérez-Rosas, and Mihai Burzo. 2013. Automatic detection of deceit in verbal communication. pages 131–134. ACM Press.

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. Association for Computational Linguistics.

Travis E. Oliphant. 2006. *A Guide to NumP*. Trelgol Publishing.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501.

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-Volume 1*, pages 309–319. Association for Computational Linguistics.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 12:2825–2830.

James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. The Development and Psychometric Properties of LIWC. page 22.

Veronica Perez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Y. Xiao, C.J. Linton, and Mihai Burzo. 2015. Verbal and Nonverbal Clues for Real-life Deception Detection. *EMNLP*, pages 2336–2346.

Veronica Perez-Rosas, Rada Mihalcea, Alexis Narvaez, and Mihai Burzo. 2014. A Multimodal Dataset for Deception Detection. *LREC*, pages 3118–3122.

Taylan Sen, Md Kamrul Hasan, Zach Teicher, and Mohammed Ehsan Hoque. 2018. Automated Dyadic Data Recorder (ADDR) Framework and Analysis of Facial Cues in Deceptive Communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22.

Yuiko Tsunomori, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. An analysis towards dialogue-based deception detection. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 177–187. Springer.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A Professional Framework for Multimodality Research. page 4.

Xiang Yu, Shaoting Zhang, Zhennan Yan, Fei Yang, Junzhou Huang, Norah E Dunbar, Matthew L Jensen, Judee K Burgoon, and Dimitris N Metaxas. 2015. Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. *IEEE transactions on cybernetics*, 45(3):492–506.