# Would I Lie To You: A Multimodal Dataset for Deception Detection

Matianyu Zang*, Tahiya Chowdhury†, and Amanda Stent‡ Davis institute for AI, Colby College
Maine, USA
Email: *matianyu_zang@brown.edu, †tahiya.chowdhury@colby.edu, ‡ajstent@colby.edu

***Abstract*—The act of deception occurs often occurs during everyday conversations in a group setting. To date, most research has either focused on laboratory settings for dialogue between two parties, and there has been very little comparison of different methods. We introduce a multimodal dataset containing multi-party deceptive and truthful conversation in a public environment during the British TV show 'Would I Lie to You', where participants compete in teams to guess truthful and deceptive statements of others. We curate the speech events from publicly available videos of the show via speaker segmentation and speech recognition. Using verbal and non-verbal behaviors of the speakers through acoustic, linguistic, and visual features, we explored classification for deception detection. Our experiments show high-performing classification results using verbal only, non-verbal only, and using both when compared to the human baseline. This work has implications for real-world deception in multi-party conversations and can inform about cues used by humans for trust and deception in multi-party settings.**

***Index Terms*—Deception detection, multi-modal learning, speech, language, affective computing.**

## I. Introduction

Deception Detection is the act of determining whether a person is being truthful or deceptive. Detecting deception can be an important skill in personal life and is an active research area for its promise of usefulness in investigative scenarios where trust is warranted, such as court trials, law, intelligence, business, etc. Research suggests that lying in high stake situations can leak observable behavioral change, that can be used to detect deception [1], [2]. However, human performance at deception detection without training is only slightly better than random chance (54%) [3]. At the same time, human confidence in their ability at the task is negatively correlated with their true ability, suggesting a false perception of their ability [4]. As a result, automated deception detection has garnered much attention from researchers, where spoken language cues are used to automatically detect deceptive speech. In recent years, research communities have explored deception detection using multimodal data including verbal cues from acoustic-prosodic and linguistic information [5], [6]) and non-verbal cues such as facial expression [7] and gesture [8]. The importance of different modalities' access in human and machine's ability to detect deception has been explored for automated deception detection from videos as well in an end-to-end manner [9], [10].

For building an automated deception detection model, an annotated multimodal dataset is necessary. [11] introduced a multimodal dataset of real deception during court trials, where trial video clips were annotated for verbal (linguistic) and non-verbal behaviors (hand gesture, gaze, and facial expression) in relation to deception. Another recent popular dataset depicting deceptive dialogue is Box of Lies [12], a multi-modal dataset containing deceptive conversations between participants playing The Tonight Show, hosted by Jimmy Fallon, Box of Lies game, in which they try to guess whether an object description provided by their opponent is deceptive or not. While other datasets can be found that further include physiological measures such as EEG signals as well as verbal and non-verbal measures [13], they are usually collected in controlled laboratory setup and often require sophisticated equipments. For detecting deceptive behavior across language and culture, the Columbia X-Cultural Deception Corpus [14] has been introduced which includes deceptive interviews in English and Chinese. In most cases, datasets are annotated by human crowd-workers which provides a human detection baseline. While these datasets are useful in deception detection research, their usefulness in detecting real-life deceptive statements can be limited for two reasons: either 1) due to reliance on manual annotation of features, and 2) unavailability of in-situ measure of human ability for deception detection, 3) lack of multi-party scenario. There is a need for a general-purpose dataset in deception detection literature that provides conversational partners' guesses about deception to serve as the human baseline.

To address this, in this paper, we describe a dataset of public videos with audio, text, and visual modalities designed for building multi-modal deception detection systems. We present a novel dataset consisting of 125 public videos that include multiple deceptive and truthful statements. The videos are collected from the TV game show 'Would I Lie You' where celebrities compete against each other to state unusual facts about themselves to deceive the opponent. While prior work has explored datasets collected from TV shows depicting deception in games played between celebrities, deceptive statements in our data occur in a multiparty conversation setting which can be useful for deception detection in group scenarios. We provide manual annotations of deceptive and truthful statements from the episode videos with samples containing rich information about the speaker's verbal and non-verbal information during the speaking event. While prior similar work has relied on manually derived linguistic and visual cues to describe the speaker events with multi-modality, we refrain from such hand-annotated features to eradicate time-consuming, subjective manual annotation. Instead, we rely on fully automatic extraction of modality-specific features leveraging existing methodologies for audio, video, and text analysis. Such an automatic approach reduces the time and cost involved in annotation, improve scalability, and practical use for deception detection. To demonstrate the performance, we benchmark our approach by evaluating its performance for training learning models proposed in prior literature in both unimodal and multimodal conditions.

The contribution of this work is the following:

- We present a novel dataset consisting of 125 public videos that contain deceptive and truthful statements in a multi-party conversation setting.
- We leverage methods developed in Natural Language Processing, Speech, and Computer Vision to extract linguistic, acoustic-prosodic, and visual features to obtain multimodal information that can aid to distinguish between deceptive and truthful speech.
- We evaluate the approach by building several models from existing literature using verbal and non-verbal information including recently introduced transformer-based models. Our experiments show that the best multimodal model provides deception detection results with a 0.96 F1 score, which is significantly above chance level and human performance.

## II. Related Work

### A. Deception detection datasets

To date, researchers resorted to various datasets for deception detection studies using different modalities of information. A wide range of studies has relied on text analysis for detecting deception from online written content. Works along this have explored travel reviews [15], dating websites [16], consumer feedback [17], and Twitter trolls [18], to identify fake and spam content. [19] created a dataset for deception detection in text containing 800 fake and online reviews collected via crowdsourcing. Such works have leveraged linguistic cues such as the use of positive and negative words, number of words, sentences, affective markers, etc. to differentiate between truthful and deceptive written statements [20].

Early research in deception detection was inspired by the polygraph, a device designed to detect lies from physiological changes such as blood pressure, pulse rate, skin temperature, and breathing [21]. Along with this idea, deception detection using EEG signals capturing signatures of brain activities has been studied and several datasets using EEG have been introduced [13], [22]. However, due to strong dependence on wearable equipment (e.g. EEG headsets), such methods have limited 'ecological validity' in real-world scenarios. Additionally, there are prior works that suggest relying on only physiological measures gathered in the laboratory setting can provide biased and misleading outcomes [23].

Based on promising early research on deception detection with micro-expression, gesture, and other non-verbal behaviors [30], [31], many studies have utilized different behavioral cues exhibited by the speaker during deceptive speech. [32], [33] utilized hand and facial features for deception detection from such a behavioral approach. [34], [35] found a decrease in iconic and deictic gestures, and an increase in metaphoric gestures to be associated with deceptive statements. Along this line, [36] showed that increased speech prompting gestures and rhythmic pulsing gestures were associated with truthful behavior.

Focusing on the interaction between the people involved during deceptive dialogues, [37] explored the nature of the conversation and the behavioral response involving strong deception resulting from it. In a similar experiment on facial expressions during deceptive conversations, [38] showed that interlocutors exhibited different facial expressions when they were told a lie as opposed to when they were told the truth. In a different study on non-verbal behavior, [39] measured interactional synchrony in head movements and facial expressions between conversation partners to differentiate between deception and truthful statements. [6] examined the influence of linguistic, gender, and native language on deception detection during interviews.

Understanding the success of individual modalities in deception detection, recent works have focused on combining different modalities towards deception detection using multimodal information. [40]–[43] focuses on multimodal deception detection using acoustic, prosodic, lexical, and visual information extracted manually or automatically for classification goals. To support this line of a multimodal approach to the problem, a collection of multimodal datasets for detecting deception from acoustic-prosodic, physiological, facial, and lexical features have been collected and released. An early data set on deceptive speech Columbia/SRI/Colorado (CSC) Corpus [40] contains 32 hours of speech collected from participants through financial incentives to deceive or tell the truth with financial reward. [28] presented a similar dataset with deceptive statements on open-domain topics collected via Amazon Mechanical Turk. [26] and [29] are two multimodal datasets for deception detection that contains both physiological and audio-visual information. Bag of Lies [13] is a recent dataset that includes EEG and gaze information along with audio and video.

However, the aforementioned datasets are generally collected via controlled experiments in laboratory settings. Box of Lies dataset [25] is a dataset collected to address this by capturing deception occurring during deceptive dialogues between a contestant and a game show host, which is close to the setting of our dataset. Real Life Trials Dataset [24] is another dataset collected by the same authors containing deceptive and truth statements made by defendants and witnesses in real-world trials. However, while deception examples collected under high-stake and multi-party conversation settings like this are informative, such real-world data and labels collected based on trial outcomes and police verification can be hard to obtain due to ethical and privacy considerations.

In this work, we consider deception detection in out-of-laboratory condition and introduce a dataset on multi-modal deception detection in multi-party dialogue settings.

### B. Features used in deception detection

Previous research studied deception detection from different perspectives using various modalities. In this section, we

| Dataset | Modalities | Speakers | Total | Dialogue type | Setting |
|---|---|---|---|---|---|
| Real Life Court Trials [24] | Video, Audio, Text | 56 | 121 | Multi-party | Real world |
| Box of Lies [25] | Video, Audio, Text | 26 | 1056 | Two-party | Real world |
| Bag of Lies [13] | Video, Audio, Gaze, EEG | 35 | 325 | - | Laboratory |
| Multimodal Deception [26] | Physiological, Thermal, and Video | 30 | 150 | Two party | Laboratory |
| CSC Corpus [27] | Audio, Text | 32 | - | Two-party | Laboratory |
| Open Domain Set [28] | Text | 512 | 7168 | - | Crowd-sourced |
| DDPM [29] | Audio, Video (RGB, Thermal), heart rate | 70 | 1680 | Two-party | Laboratory |
| **Would I lie to You (WILTY)** | Audio, Text, Video | 245 | 502 | Multi-party | Real world |

TABLE I
AVAILABLE MULTIMODAL DATASETS FOR DECEPTION DETECTION.

discuss the three categories related to this work: linguistic, acoustic-prosodic, and visual-physiological.

**Linguistic.** This category refers to features extracted from the spoken text during deceptive statements. This includes n-gram, part of speech tag [19], [44], word counts per review or utterance, complexity (average word length), diversity (number of unique words over the total number of words) [15], verb choice in statements [45] non-immediacy (self-reference, group-reference, generalization, indefinite articles), uncertainty (modifier, modality, subjectivity, quotation, question, hedge words) [18], syntactic style [46], and specificity (disclosure markers, causation, sense terms, use of numbers, relativity). In addition to those, Linguistic Inquiry and Word Count (LIWC) is a frequently used text analysis software resource [6] that classifies words into 80 different linguistic, psychological, or topical categories [47]. Word Embeddings [48], bag of words [41], [43] are also used to obtain informative feature representation from dialogue and statement transcripts for deception detection.

**Acoustic-prosodic.** Acoustic-prosodic cues extracted from speech have been used to characterize deceptive speech in prior works. One such frequently used indicator is Mel-frequency Cepstral Coefficients (MFCC) [49], which transforms audio signals into frequencies for improved speech recognition and other speech tasks. [5] used several commonly used speech features: intensity, pitch, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate for detecting deception, and found increased pitch and slower speaking rate to be indicative of deception. In a different study, [42] explored speech disfluency, filled pauses, response latency, false starts, and repetition of words as indicators for detecting deception and found them used as useful deception indicators by humans despite their unreliability. Like LIWC used for linguistic feature analysis, prior works have used speech analysis tools such as Praat [50] to obtain statistical components of acoustic and prosodic information such as shimmer, voice quality, loudness, pitch, duration, and speaking rate as possible indicators of deception. In several more recent works, openSMILE, an audio feature extraction toolbox [51] is frequently used to extract Interspeech2013 (IS2013) ComParE Challenge baseline feature set [52] for deception detection, including pitch (fundamental frequency), intensity (energy), spectral, cepstral (MFCC), duration, voice quality (jitter, shimmer, and harmonics-to-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness.

**Visual-physiological.** Early research on micro-expressions as indicators for detecting a lie [30], [53] inspired use of Facial Actions Units towards deception detection [26], [43]. [43] also used different face encoding techniques such as Fisher Vectors to automatically extract visual features from faces during deceptive statements. Among other non-verbal behaviors, [34]–[36] utilized speaker gestures (iconic, deictic) to differentiate between deceptive and truthful behavior. To capture physiological activity change during deceit, EEG signal [13], brain imaging [54], pulse rate [29], skin temperature and conductance [26] have been used in prior works. To understand the association between eye contact and deception, eye gaze [13] and head movement [39] have been found useful.

Fig. 1. Snapshot of a card reading segment from 'Would I lie to you' season 4 episode 4.

Apart from linguistic, acoustic-prosodic, and visual features, prior works have explored speakers' identity and personality-specific traits such as age, gender [28], native language, and five-factor personality inventory score (extroversion, agreeableness, openness, conscientiousness, and neuroticism) [41], [5] for deception detection.

## III. DATA

To explore automated deception detection during dialogue, we collect data from conversations of deceptive behavior. Inspired by prior deception detection dataset collected from guessing game show, we opted for a game suitable to this setting.

### A. Data Source

We use episodes of the British TV series 'Would I Lie To You' as our source of data. Each episode in this series features

| Speaker | #Truth | #Lie | Truth | Lie |
|---|---|---|---|---|
| Host Speaker | 84 | 50 | I once simultaneously worked as both the DJ and the newsreader on local radio using a different accent for each job. | I believe disaster will occur if I don't adhere to my special alarm clock system. |
| Regular Speaker | 173 | 213 | I have no idea how to use a launderette washing machine. | I was a teenage boxer. But I quit because I had a mouth ulcer. |

TABLE II
EXAMPLE OF UTTERANCES IN THE DATASET.

four guest public figures who come from different backgrounds including acting, comedy, sports players, newscasters, politicians, religious leaders, etc. For each episode, the guests join one of two teams each headed by one of the two co-hosts of the show. The two teams then compete each other where each player reveals facts and personal tales for the consideration of the members of the opposing team. The opposite team's task is to separate the true facts from the fabricated lies.To explore automated deception detection during dialogue, we collect data from conversations of deceptive behavior. Inspired by the prior deception detection dataset collected from a game show, we opted for a TV show suitable to this setting.

### B. Dataset Collection

Each episode runs for 28 minutes on average with several rounds each with its unique rule. Episodes have three main sections: a long-segment section, a guess-who-it-is section, and a short-segment (quick-fire-lie) section. In each segment in the long- and short-segment sections, one of the six participants is 'randomly' selected to turn over a card on their desk. This card describes an event, or fact or occasionally refers to an object, a possession that relates to the speaker. The participant must describe the event convincingly as if it happened to them or the object belongs to them, and the other team must guess whether the participant is telling the truth or a lie. After some minutes of Q&A, the participant reveals whether they have been telling the truth or telling a lie by pressing the button in front of them. In the guess-who-it-is section, a visitor, who has a genuine connection with one of the players, is invited to the stage. All three members of the team describe their relationship and encounter with the visitor. It is left for the opposing team to guess who among the players has a genuine relationship with the guest.

For this research, we only extract the information from the long and short segment sections. More specifically, we only clip the snippets spanning from each speaker picking up the card to finish reading the statement. We do not include the Q&A interactions that follow in the data pre-processing steps. We collected 125 videos from the 14 seasons of the series running between 2007 to 2021. Note that some of the guests have reappeared in the show and the two hosts are consistent in each episode, resulting in multiple rounds from some of the players in the data.

Before we describe the details of pre-processing, we define the following terms for the readers:
**Section**: It refers to about 2-10 minute long fragments of the video. In each section, a complete round of the game and player interactions (statement declaring, question answering, guessing, and result revealing) take place.
**Utterance**: Those terms refer to the card reading event, usually lasting for only seconds, during which speakers read the statement out loud to the other players of the show.

### C. Speaker Segmentation

- We extract the audio from each video, downsample it to 16kHz, and reduce it from a dual channel into a mono-channel .wav files to ensure lossless audio quality.
- We automatically segment the episode into sections using some signal sounds. There are four types of them. 'Speaker-indicator beep' and long silence (when the speakers pick up their card) are used for detecting the start of each section, while 'answer-revealing sound' delimits the end of each section. Meanwhile, long silence is also used to skip the title sequence, while similarly the end-of-episode buzz signals the end of the last section, skipping the closing credits. These signals together help segment the episode into multiple sections.
- We use diart [55] to perform speaker diarization, which takes audio of the segment and identifies the active speaker of each spoken sentence. This allows us to get the first speaker turn in each statement section. The first speaker turns are the utterances that speakers read the statements on cards.
- With the preliminary segmentation, we manually verified the start and end time of each section and calibrate the first speaker turn boundary in each section to capture the full utterance.
- We use Prodigy [56] annotation software to label the speaker identity as well as the true status of the statement (Truth or Lie). Prodigy provides a multimodal annotation platform on which regions in an audio-visual recording can be annotated with start, and end times along with labels for the regions, which can be a section or an utterance. This region is later used to obtain features from the audio, transcription, and video file of the utterance.

### D. Transcription

After the segment boundaries and labels are obtained through annotation, we use the start and end time stamps of each utterance segment to get utterance audio as a .wav file. We use wav2vec 2.0 [57] to perform speech recognition from the audio and get the transcript of each speaker's utterance delivery. Note that we observe no case of overlapped speech (during card reading of a TV show) in our data, which may not be the case for all deceptive conversations. Table III-B shows example transcripts for both Truth and Lie classes in the dataset.

## IV. Methodology

We now describe our methods for extracting features from audio, text, and video modalities from the utterance samples of the dataset.

### *A. Acoustic-prosodic Features*

Prior works [41], [43] have identified several acoustic prosodic features useful for deception detection. We used the Interspeech 2013 (IS13) ComParE Challenge baseline feature set from openSMILE, a standard feature set for deception detection and many other computational paralinguistic tasks, which resulted in about 6000 features in total per sample. Since males and females have different voice attributes in nature, we normalize the acoustic features by gender.
In addition, according to [42]'s work, which indicates a close connection between filled pauses and deceptive utterances, we detect filled pauses, the presence of filled pauses in the utterance (as a boolean variable, and the number of filled pauses extracted via Praat. Finally, we also extract the duration of the single utterance and the whole section that contains the corresponding utterance as a feature.

### *B. Linguistic Features*

To replicate the selection of linguistic features (lexical, syntactic, semantic, pragmatic) features used in prior work on this problem, we refer to [42], where the authors did p tests on a wide range of lexical/acoustic/sentimental features. We choose only features with a significance level less than 0.001, i.e. features that are tested to be strong indicators of deception/truth. In addition, we also include N-grams and Word Embeddings representations of the transcripts as leveraged in previous research like [41].

**Pragmatic.** We include four pragmatic features: presence and number of hedge and weasel words. Weasel words usually make sentences uncertain or hollow (e.g. 'many', 'really', 'seemed'), while hedge words convey ambiguity and indecisiveness (e.g. 'possibly', 'a few', 'something'). Truthful statements are more concrete and specific in content, so weasel words and hedge words are considered good indicators of deceptive statements.

**Lexico-syntactic.** We count the number of sentences per statement, words per statement, words per sentence, and the number of words that is at least six-character long. Furthermore, to understand the sentence structure, we count the verbs, nouns, adjectives, and numbers in each statement. We also employ N-gram features for $n = 1, 2, 3$ using the binary counting system. We enumerate all unique unigrams, bigrams, and trigrams in the datasets and calculate their frequencies. We keep only N-grams that appear at least five times across all statements.

**Semantics.** As for semantics, we pair each word with the GloVe embeddings for inter-word semantic relationships. We use the 200-dimensional model trained on 2 billion tweets. Also, we include the concreteness score of each word in the statement.

### *C. Visual Features*

To extract visual features from the video segment, we downsampled the video to 10 frames/second. Following prior work [43], we encode the faces of the speakers during utterance to automatically extract visual features. To extract visual features from each frame, we use deepface [58] library which hosts a collection of pre-trained face recognition and embedding models. DeepFace library identifies the faces present in the image and represents each face image as a $n$-dimensional vector embedding through pre-trained facial representation models, where $n$ varies between 128 to 4096 depending on the model implementation. We have experimented with VGG-FACE [59], FaceNet [60], OpenFace [61], SFace [62] and DeepFace [63] to provide fixed size embedding per image. We compute the mean embedding vector for the video segment during an utterance by averaging across all frames of the samples, which we use as the visual features for each sample.

Since the two hosts (Rob and Angus) and co-hosts (David and Lee) present frequently on the show, our dataset included a significantly higher number of samples uttered by each of them compared to the guests. To avoid over-fitting resulting from this, we exclude their statements from the original data set. This results in two smaller data sets, one for hosts only and one for all visiting guests (regular-speaker). The final host-only data set has 136 samples, and the regular-speaker data set has 365 samples each represented by 6000 features.

### *D. Multi-modal Feature Fusion with Transformers*

Prior works have explored multimodal feature extraction from videos for deception detection purposes [10]. Instead of extracting acoustic-prosodic, lexical, and visual features through a modality-specific framework, we looked for an automated modality-agnostic feature extraction method that can leverage the three input modalities (text, audio, and images) present in the video in a general architecture.

**Multi-modal Transformers Architecture.** Because of this specific property, we choose Perceiver IO [64], a general-purpose transformer-based architecture designed for handling data from multiple modalities and multi-modal tasks with no changes required to the architecture. Specifically, Perceiver IO uses attention to map inputs of a wide range of modalities to a fixed-size latent space that is further processed by a deep, fully attentional network. It further offers a mechanism to decode the latent space to flexibly produce outputs using a querying method specific to a range of domains and can then be used for classification tasks.

In this work, we use the Perceiver Model implementation available at HuggingFace [65]. While Huggingface provides pre-trained models for text, image, and audio, direct finetuning on our dataset was unsuccessful for two reasons: due to the small sample size of our dataset (365 samples) compared to the 250 million parameters of the model, and the data used in the pre-training was not well suited to our goal of deception detection. Instead, our strategy was to pre-train a multi-modal perceiver model on a larger dataset created for a similar task.

Towards this goal, we choose the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset

[66], a multimodal dataset for emotion recognition from facial, textual, and audio data. The data contains video and audio recordings of 24 professional actors (12 female, 12 male) stating a set of statements in 8 emotion expression conditions: calm, happy, sad, angry, fearful, surprise, disgust, and neutral. The dataset contains 1440 pairs of audio and video files, and corresponding statement text which we used as the transcript of the speech events.

**Pre-training.** For pretraining a multimodal perceiver model, we followed the setup proposed in [67], which introduces a multimodal perceiver-based architecture that combines video frames, audio, and text for emotion recognition task at MuSe '22. In this architecture, Vision Transformer (ViT) [68], RoBerta contextualized embeddings [69] and WavLM [70] is used to produce latent features from video, text, and audio samples of the RAVDESS dataset. We combine these three sets of latent features to train a Perceiver Classification model using early stopping criteria for 50 epochs that we test on a hold-out test set.

To finetune this pre-trained model on our dataset, we use the same train and test split as used in our other experiments. Our pre-trained model is finetuned for a classification setting of truth and lies classes for 15 epochs.

## V. Experiments

We now present our modeling approaches using the features extracted as described above. We experiment with various machine learning modeling approaches that prior works [5], [41], [43] have used for deception detection tasks. Since Lee and David present on almost every episode of the show and thus have unproportionately many records, we train models on regular speakers' data to ensure no speaker has more than ten statements in the data set.

For training the classification model in the supervised setting, we use the true identity of the statement (Truth, Lie) as the ground truth label. We randomly split all the records into training and testing sets so that 80% of the data is used for training and the rest for testing. Since some speakers had multiple statements, we ensured speakers in the training set do not appear in the testing set when splitting the data to ensure model robustness. We repeat each experiment 10 times and report the average of the performance metric. We also preserve the same ratio between truth and lie classes in training and test sets to have a balanced dataset. We use standard performance measures for classification problems such as accuracy, precision, recall, and F1-score as our evaluation metric. Our models and experimental pipelines are implemented using scikit-learn, Keras, and PyTorch libraries.

We first conducted experiments with verbal and non-verbal information separately (video, text, audio), and then combined the modalities. We first explored the following experimental settings: Logistic Regression, Decision Tree, Gaussian Naive Bayes, Multi-layer Perceptron, and Support Vector Machine with a linear kernel. We also included two ensemble methods: Random Forest and Adaboost for classification. All models used the default parameters in scikit-learn with max iteration set to 2000. To compare with prior work [14], we also trained Bidirectional LSTM with one Bi-LSTM layer of 256-dimensional input. Our batch size was set to 32 with Adam Optimizer. Understanding our data set has several thousands of features, we also conduct Principal Component Analysis (PCA) to see if dimensionality reduction can improve the result.

| Name | Accuracy | Count |
|---|---|---|
| David Mitchell | 59.22 | 179 |
| Lee Mack | 59.64 | 166 |
| Jason Manford | 36.36 | 11 |
| Claudia Winkleman | 54.55 | 11 |
| Jimmy Carr | 44.44 | 9 |
| Bob Mortimer | 44.44 | 9 |
| Gabby Logan | 62.50 | 8 |
| Jo Brand | 50.00 | 8 |
| Richard Osman | 62.50 | 8 |

TABLE III
Lie detection accuracy for speakers with no fewer than eight attempts

For the multimodal model, we used the same settings while using the features combined from all three modalities through the early fusion technique. In the multimodal model with finetuned transformer model, we used the multimodal perceiver implementation described in [67] with a batch size of 16 and a learning rate of $1e^{-}5$. All experiments are done on 8 GB GPU memory except finetuning transformer which used 24 GB GPU memory.

## VI. Results

### A. Human Performance Baseline

Previous studies have showcased that humans are notoriously poor at lie detection, achieving success rates only slightly better than random guessing, 54% [71] and 56.75% [5]. To understand how well human detects lies in the scenario of our dataset, we manually label all speakers' guesses of statements. This becomes our human baseline for deception detection. In each turn, three members of the opposite team attempt to guess whether the speaker's statement was true or a lie, and the team captain (Lee or David) announces the final decision. In many cases, contestants withhold their decision, so there are three or fewer guesses available for each statement. In the end, we collect 999 guesses, 762 from male contestants and 237 from female contestants. In general, Overall deception detection performance has an accuracy of 55.86%, with males (56.69%) doing a slightly better job than females (53.16%), which is similar to the observation made in prior work.

Since a subset of contestants has multiple samples in our dataset, we also inspect individual performances for speakers with at least eight guesses. Table III demonstrates the guessing accuracy of contestants who have eight or more attempts.

We observe that no contestant performed significantly better at the task of deception detection than random guessing. About half of the speakers performed worse than random guessing and the rest performed slightly better. Note that while Lee and David, the two co-hosts of the show have appeared in almost all episodes for 14 years, their deception detection performance did not significantly improve from this experience

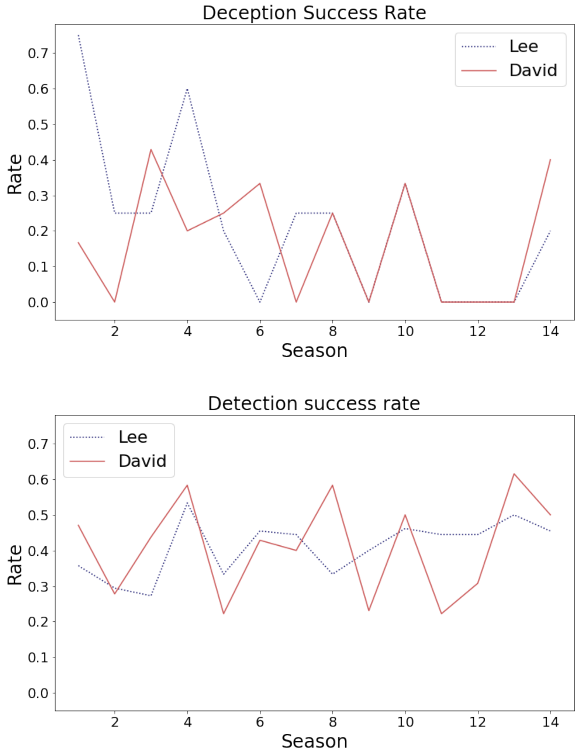

Fig. 2. Trend of deception success and deception detection for Lee and David in each season of the show.

(10 times more attempts than regular speakers). While both co-hosts are more accurate than some of the machine learning models that we present later in the paper, they perform lower than most models (see model performance in the next section).

In addition to the overall accuracy, we are also interested in whether the skill of deception and detecting deception can be trained due to practice. Given that Lee and David have appeared in each season, we expect their performance to improve over time from the experience. In Figure 2, we show the trend of successful deception and detection for the two co-hosts throughout all 14 seasons. We do not observe improved deception and detection performance with more practice as the show progresses. This is an observation observed in other multi-modal deception detection experiments using data collected from game show [25], where the host, despite getting more practice, does not perform significantly better than the guests appearing often only once in the show. One explanation for this can be training humans in lie detection can be difficult considering the various techniques the contestants, who are generally professionals working in the entertainment and acting industry, employ to deceive the opposing team, making it difficult to be skillful in deception detection. To the best of our knowledge, the game rounds in the Would I Lie to You show involve spontaneous, non-scripted interactions meaning the participants devise their own method of deception and detection during the gameplay. As Figure 2 shows, neither hosts achieve accuracy higher than 62.50% which is the best performance among all speakers who at least 8 attempts present in the dataset.

### B. Multimodal Deception Detection

**Verbal features.** Now that we have a human baseline, next, we present results from our deception detection model experiments. We first trained classification models using features from verbal behaviors: acoustic-prosodic and lexical features. Specifically, we used Logistic regression (LR), Naive Bayes, Decision Tree, Multi-layer Perceptron (MLP), Adaboost, Random Forest, Linear SVM, and Bidirectional LSTM. We compared the results of the different classification models in Table IV.

LR, Naive Bayes, MLP, Adaboost perform quite similar (62%-68% accuracy, 0.61-0.68 F1 score). However, we observe that some models have better performance for the Lie class than the Truth class (High recall rate: LR, MLP, Adaboost, Random Forest) which suggests the models perform well at detecting deception. On the other hand, we observe a high precision rate for the Truth class in LR, MLP, and Adaboost, which means these classification models are stronger in the Truth class. We find that BLSTM performed the best among all the models with 96% accuracy and 0.96 F1 score for both classes.

Since our feature set involves $> 6000$ features, we employ feature reduction using the dimensionality reduction technique and trained the same models with a reduced number of features. We apply Principal Component Analysis (PCA) for dimensionality reduction and experimented with a range of values for the number of principal components $n$. We chose $n = 300$ as that retained $> 95\%$ variance. Applying PCA improved results for some classification models (LR, Decision Tree, Linear SVM), but in other cases, performance degraded. BLSTM, which reported the best performance also degraded ($96\% \rightarrow 93\%$) specifically in the truth class which suggests reduced features as useful for the truth class.

**Non-verbal Features.** To explore the role of visual features in deception detection, we next trained classification models using visual features extracted from the video as described in section IV-C. Note that, the vector representation of the visual features has varying dimensions depending on the output of the pre-trained embedding model (VGG: 2622, FaceNet: 128, OpenFace: 128, DeepFace: 4096, SFace: 128). We applied PCA to reduce dimension while maintaining $> 95\%$ variance and present the results in Table V.

We found that representation obtained from OpenFace performs the best as visual features in deception detection task with 86% accuracy and 0.87 F1 score. Due to its significantly better performance than other visual feature embeddings, we use OpenFace for the rest of our experiments.

We compared the results of the different classification models trained with visual features in Table VI. . Here again, we find that BLSTM performed the best among all the models with 81% accuracy and 0.80 F1 score. After applying PCA, the performance of the BLSTM model further improved (($81\% \rightarrow 96\%$) ) with faster training time as the latter model involved fewer parameters.

**Multi-modal Learning for Deception Detection.** After experimenting and training models with the modalities separately, we trained multimodal learning models combining acoustic-prosodic, lexical, and visual features. We choose

| | **Before PCA** | | | | | | | **After PCA** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Truth** | | | **Lie** | | | | **Truth** | | | **Lie** | | |
| **Method** | **Accuracy** | **P** | **R** | **F1** | **P** | **R** | **F1** | **Accuracy** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| LR | 62% | 0.70 | 0.44 | 0.54 | 0.59 | 0.81 | 0.68 | **67%** | 0.65 | 0.61 | 0.63 | 0.68 | 0.71 | 0.70 |
| Naive Bayes | 65% | 0.63 | 0.75 | 0.69 | 0.67 | 0.53 | 0.59 | 55% | 0.58 | 0.48 | 0.52 | 0.52 | 0.62 | 0.57 |
| Decision Tree | 50% | 0.50 | 0.33 | 0.40 | 0.50 | 0.67 | 0.57 | 61% | 0.62 | 0.50 | 0.55 | 0.60 | 0.71 | 0.65 |
| MLP | 68% | 0.75 | 0.60 | 0.57 | 0.62 | 0.77 | 0.69 | 67% | 0.67 | 0.67 | 0.67 | 0.67 | 0.67 | 0.67 |
| Adaboost | 68% | 0.75 | 0.50 | 0.60 | 0.64 | 0.84 | 0.73 | 56% | 0.57 | 0.50 | 0.53 | 0.56 | 0.62 | 0.59 |
| Random Forest | 68% | 0.77 | 0.53 | 0.62 | 0.64 | 0.84 | 0.73 | 53% | 0.50 | 0.27 | 0.35 | 0.54 | 0.76 | 0.63 |
| Linear SVM | 69% | 0.59 | 0.87 | 0.70 | 0.71 | 0.36 | 0.48 | **71%** | 0.65 | 0.81 | 0.72 | 0.79 | 0.61 | 0.69 |
| BLSTM | **96%** | **0.97** | **0.94** | **0.96** | **0.94** | **0.97** | **0.96** | 93% | 0.97 | 0.88 | 0.92 | 0.89 | 0.97 | 0.93 |

TABLE IV
DECEPTION DETECTION CLASSIFICATION RESULT WITH AUTOMATED ACOUSTIC-LEXICAL FEATURES.

| **Method** | **Accuracy** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|
| DeepFace | 0.72 | 0.74 | 0.76 | 0.74 |
| VGGFace | 0.68 | 0.80 | 0.61 | 0.67 |
| FaceNet | 0.55 | 1.0 | 0.50 | 0.71 |
| OpenFace | **0.86** | **0.87** | **0.89** | **0.87** |
| SFace | 0.81 | 0.84 | 0.81 | 0.82 |

TABLE V
DECEPTION DETECTION CLASSIFICATION RESULT WITH AUTOMATED VISUAL FEATURES EXTRACTED USING OPENFACE LIBRARY. DUE TO THE LARGE NUMBER OF FEATURES WE APPLIED PCA TO REDUCE DIMENSION TO RETAIN $> 95\%$ VARIANCE. NUMBER OF PRINCIPAL COMPONENTS CHOSEN BETWEEN (300-100).

BLSTM for this final model as that has performed well in the previous setting (see Tables IV and VI. We also compare these models trained with modality-specific features combined via early fusion to a model trained with modality-agnostic features using a transformer-based architecture in Table VII.

Surprisingly, we found verbal (acoustic and linguistic) features to be the best-performing one with a 0.96 F1 score. The model trained with all 3 modalities' combined features (acoustic, linguistic, visual) performed slightly lower (0.87 F1 score). Perceiver, the transformer-based model, performed significantly lower with a high recall rate (0.67 F1 score). One explanation for this can be modality-specific pre-processing that was not available in the feature extraction process within the general-purpose transformer architecture. Another issue can be the small number of training examples that were used in fine-tuning Perceiver, which has a large number of parameters to tune.

## VII. DISCUSSION

Our experiments show verbal information to be more informative for deception detection than non-verbal (visual information). This is also observed in prior work on multimodal deception [11], [25] which shows acoustic-linguistic to hold strong indicators of deception. Unlike observed in these prior works, using non-verbal information (visual modality) performance is worse than using verbal features. Furthermore, adding both non-verbal and verbal information to the model worsen the performance, which is contrary to observations made in [11], [25]. One explanation for this can be that these works used manually annotated gestures (head movement, mouth, gaze, etc.) as non-verbal behaviors as opposed to our approach of using face encoding from the videos as automatically extracted visual features. Such automatic feature extraction can contain noise due to face occlusion in general because of camera angle and lighting. Due to the multi-party setting where the speaker is speaking not only to the audience but also to the other contestants, which influences the direction of eye-gaze and head movement, features which are found as useful non-verbal behaviors for classification. Table IV shows that acoustic-linguistic features are more useful in detecting truth than lies. In the future, adding dialogue features to understand the conversation cues between players during Q&A after the statement can provide more interesting insights into human behavior for deception detection.

Our experimentation with dimensionality reductions revealed interesting insights. For non-verbal features, feature reduction has degraded or maintained the performance for all learning models, whereas for verbal features performance improved in most cases after reducing features. This suggests the importance of the complete set of verbal features in classification, but a selected subset of verbal features can be sufficient for building a good classifier. While this work relies on automated visual feature extraction, exploring different gesture expressed by non-verbal behavior (body, face, hand) and their importance in classification remains for future work. Our experiment with modality-agnostic multimodal learning with Perceiver architecture indicates the effectiveness of modality-specific feature extraction and using them in modeling via early fusion. Recent works on neural networks parameter pruning using Lottery Ticket Hypothesis [72], [73] showed the existence of useful subnetworks within large Transformer-based networks with fewer trainable parameters. Exploring such subnetworks to reduce training data requirements and improve performance for modality-agnostic training can be another direction for future work.

### *A. Ethical Considerations*

The work presented in this paper is based on data collected in conditions that may not be applicable to a real-world setting (e.g. ample lighting in the studio, high-quality audio recording, etc.). The machine learning models we present here is for research purpose and should not be used without human intervention, particularly in high stake scenarios concerning public security and crucial life outcomes like interviews or trials. While we plan to release the dataset for research purposes only (available upon request), we do not plan to release the models due to ethical considerations.

| | Before PCA | | | | | | | After PCA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Truth | | | Lie | | | | Truth | | | Lie | | |
| Method | Accuracy | P | R | F1 | P | R | F1 | Accuracy | P | R | F1 | P | R | F1 |
| LR | 62% | 0.59 | 0.59 | 0.59 | 0.65 | 0.65 | 0.65 | 62% | 0.59 | 0.59 | 0.59 | 0.65 | 0.65 | 0.65 |
| Naive Bayes | 55% | 0.53 | 0.56 | 0.54 | 0.58 | 0.55 | 0.56 | 63% | 0.62 | 0.42 | 0.50 | 0.63 | 0.80 | 0.71 |
| Decision Tree | 53% | 0.46 | 0.38 | 0.41 | 0.57 | 0.65 | 0.60 | 54% | 0.54 | 0.37 | 0.44 | 0.54 | 0.70 | 0.61 |
| MLP | 60% | 0.59 | 0.59 | 0.59 | 0.61 | 0.61 | 0.61 | 56% | 0.53 | 0.59 | 0.56 | 0.59 | 0.53 | 0.56 |
| Adaboost | 41% | 0.36 | 0.25 | 0.29 | 0.44 | 0.57 | 0.50 | 61% | 0.61 | 0.55 | 0.58 | 0.61 | 0.67 | 0.64 |
| Random Forest | 54% | 0.57 | 0.29 | 0.38 | 0.52 | 0.79 | 0.63 | 53% | 0.57 | 0.25 | 0.35 | 0.52 | 0.81 | 0.63 |
| Linear SVM | 56% | 0.70 | 0.37 | 0.48 | 0.50 | 0.80 | 0.62 | 51% | 0.46 | 0.32 | 0.37 | 0.54 | 0.68 | 0.60 |
| BLSTM | 81% | 0.83 | 0.70 | 0.76 | 0.79 | 0.89 | 0.84 | **96%** | **0.95** | **0.97** | **0.96** | **0.97** | **0.95** | **0.96** |

TABLE VI
DECEPTION DETECTION CLASSIFICATION RESULT WITH AUTOMATED VISUAL FEATURES EXTRACTED USING DEEPFACE LIBRARY.

| Method | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| Acoustic + Linguistic | 96% | 0.96 | 0.96 | 0.96 |
| Acoustic + Linguistic + Visual | 86% | 0.86 | 0.88 | 0.87 |
| Transformer (Perceiver) | 52% | 0.52 | 0.94 | 0.67 |
| Human Baseline | 62% | 0.63 | 0.62 | 0.61 |

TABLE VII
COMPARISON OF MULTIMODAL DECEPTION DETECTION PERFORMANCE WITH HUMAN BASELINE.

## VIII. CONCLUSION

In this paper, we present a dataset of publicly available videos for multimodal deception detection. Our dataset consists of deceptive and truthful statement utterances recorded in multi-party conversation settings. We considered multimodal features obtained from verbal and non-verbal information available from speech events via automated acoustic, linguistic, and visual feature processing. We established a human baseline and explored human skill development for deception detection through repeated practice. By integrating different modalities, we investigated a classification model for detecting deception. Our best classifier shows improved detection performance for both Truth and Lie compared to human baseline and random guessing with a 96% F1 score. We found verbal features to perform better than non-verbal features. SE believe curation and evaluation of such datasets and modeling approach can open up novel research opportunities to understand deception in multi-party conversation in public settings.

## REFERENCES

[1] L. Ten Brinke, S. Porter, and A. Baker, "Darwin the detective: Observable facial muscle contractions reveal emotional high-stakes lies," *Evolution and Human Behavior*, vol. 33, no. 4, pp. 411–416, 2012.

[2] L. Ten Brinke and S. Porter, "Cry me a river: identifying the behavioral consequences of extremely high-stakes interpersonal deception." *Law and Human Behavior*, vol. 36, no. 6, p. 469, 2012.

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "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*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 309–319. [Online]. Available: https://aclanthology.org/P11-1032

[4] B. M. DePaulo, K. Charlton, H. Cooper, J. J. Lindsay, and L. Muhlenbruck, "The accuracy-confidence correlation in the detection of deception," *Personality and Social Psychology Review*, vol. 1, no. 4, pp. 346–357, 1997.

[5] A. M. Sarah Ita Levitan and J. Hirschberg, "Linguistic cues to deception and perceived deception in interview dialogues," *Association for Computational Linguistics*, vol. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1941–1950, Jun. 2018. [Online]. Available: https://aclanthology.org/N18-1176

[6] S. I. Levitan, A. Maredia, and J. Hirschberg, "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)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1941–1950. [Online]. Available: https://aclanthology.org/N18-1176

[7] X. Shen, G. Fan, C. Niu, and Z. Chen, "Catching a liar through facial expression of fear," *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.675097

[8] D. Avola, L. Cinque, M. De Marsico, A. Fagioli, and G. L. Foresti, "Lietome: Preliminary study on hand gestures for deception detection via fisher-lstm," *Pattern Recognition Letters*, vol. 138, pp. 455–461, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520303123

[9] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, "Face-focused cross-stream network for deception detection in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7802–7811.

[10] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[11] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 59–66.

[12] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[13] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 83–90.

[14] K.-Z. L. Gideon Mendels, Sarah Ita Levitan and J. Hirschberg, "Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection," in *Proc. Interspeech 2017*, 2017, pp. 1472–1476.

[15] U. G. Kyung-Hyan Yoo, "Comparison of deceptive and truthful travel reviews," *Information and Communication Technologies in Tourism 2009*, pp. 37–47, 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-211-93971-0_4

[16] J. T. H. Catalina L. Toma, "Reading between the lines: Linguistic cues to deception in online dating profiles," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010.

[17] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1566–1576.

[18] A. Addawood, A. Badawy, K. Lerman, and E. Ferrara, "Linguistic cues to deception: Identifying political trolls on social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, no. 01, pp. 15–25, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3205

[19] C. C. Myle Ott, Yejin Choi and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," 2011. [Online]. Available: https://arxiv.org/abs/1107.4557

[20] J. J. F. Nunamaker, J. K. Burgoon, T. Qin, and J. P. Blair, "Modality effects in deception detection and applications in automatic-

deception-detection," in *2014 47th Hawaii International Conference on System Sciences*, vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, jan 2005, p. 23b. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/HICSS.2005.436

[21] J. Synnott, D. Dietzel, and M. Ioannou, "A review of the polygraph: history, methodology and current status," *Crime Psychology Review*, vol. 1, no. 1, pp. 59–83, 2015. [Online]. Available: https://doi.org/10.1080/23744006.2015.1060080

[22] A. Turnip, M. F. Amri, H. Fakhrurroja, A. I. Simbolon, M. A. Suhendra, and D. E. Kusumandari, "Deception detection of eeg-p300 component classified by svm method," in *Proceedings of the 6th international conference on software and computer applications*, 2017, pp. 299–303.

[23] M. Derksen, "Control and resistance in the psychology of lying," *Theory & Psychology*, vol. 22, no. 2, pp. 196–212, 2012.

[24] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 59–66. [Online]. Available: https://doi.org/10.1145/2818346.2820758

[25] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1768–1777. [Online]. Available: https://aclanthology.org/N19-1175

[26] V. Pérez-Rosas, R. Mihalcea, A. Narvaez, and M. Burzo, "A multimodal dataset for deception detection," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3118–3122. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/869_Paper.pdf

[27] J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Hoffman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," 09 2005, pp. 1833–1836.

[28] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1120–1125. [Online]. Available: https://aclanthology.org/D15-1133

[29] N. Vance, J. Speth, S. Khan, A. Czajka, K. W. Bowyer, D. Wright, and P. Flynn, "Deception detection and remote physiological monitoring: A dataset and baseline experimental results," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 4, pp. 522–532, Oct 2022.

[30] P. Ekman, M. O'Sullivan, and M. G. Frank, "A few can catch a liar," *Psychological Science*, vol. 10, no. 3, pp. 263–266, 1999. [Online]. Available: https://doi.org/10.1111/1467-9280.00147

[31] M. G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies." *Journal of personality and social psychology*, vol. 72 6, pp. 1429–39, 1997.

[32] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 36–43, 2005.

[33] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju, "Real-time automatic deceit detection from involuntary facial expressions," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007.

[34] D. Cohen, G. Beattie, and H. Shovelton, "Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed," 2010.

[35] L. Caso, F. Maricchiolo, M. Bonaiuto, A. Vrij, and S. Mann, "The impact of deception and suspicion on different hand movements," *Journal of Nonverbal behavior*, vol. 30, pp. 1–19, 2006.

[36] J. Hillman, A. Vrij, and S. Mann, "Um… they were wearing…: The effect of deception on specific hand gestures," *Legal and Criminological Psychology*, vol. 17, no. 2, pp. 336–345, 2012.

[37] Y. Tsunomori, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "An analysis towards dialogue-based deception detection," *Natural Language Dialog Systems and Intelligent Assistants*, pp. 177–187, 2015.

[38] T. Sen, M. K. Hasan, Z. Teicher, and M. E. Hoque, "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*, vol. 1, no. 4, pp. 1–22, 2018.

[39] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. E. Dunbar, M. L. Jensen, J. K. Burgoon, and D. N. Metaxas, "Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 492–506, 2015.

[40] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis *et al.*, "Distinguishing deceptive from non-deceptive speech," 2005.

[41] K.-Z. L. Gideon Mendels, Sarah Ita Levitan and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," *INTERSPEECH*, Aug. 2017. [Online]. Available: https://www.researchgate.net/publication/319185334

[42] X. L. Chen, S. I. Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 199–214, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.14

[43] J. Zhang, S. I. Levitan, and J. Hirschberg, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features." in *INTERSPEECH*, 2020, pp. 359–363.

[44] M. P. Tommaso Fornaciari, "Automatic deception detection in italian court cases," *Artificial intelligence and law*, vol. 21, no. 3, pp. 303–340, Feb. 2013. [Online]. Available: https://link.springer.com/article/10.1007/s10506-013-9140-4

[45] S. H. Adams, "Statement analysis: What do suspects' words really reveal," *FBI L. Enforcement Bull.*, vol. 65, p. 12, 1996.

[46] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 171–175. [Online]. Available: https://aclanthology.org/P12-2034

[47] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[48] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[49] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[50] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.

[51] *Opensmile: the munich versatile and fast open-source audio feature extractor*, 2010. [Online]. Available: https://dl.acm.org/doi/proceedings/10.1145/1873951

[52] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[53] P. Ekman, "Darwin, deception, and facial expression," *Annals of the new York Academy of sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[54] E. H. Meijer and B. Verschuere, "Deception detection based on neuroimaging: Better than the polygraph?" *Journal of Forensic Radiology and Imaging*, vol. 8, pp. 17–21, 2017.

[55] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1139–1146.

[56] I. Montani and M. Honnibal, "Prodigy: A new annotation tool for radically efficient machine teaching," *Artificial Intelligence*, vol. to appear, 2018.

[57] A. M. Alexei Baevski, Henry Zhou and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[58] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27. [Online]. Available: https://doi.org/10.1109/ASYU50717.2020.9259802

[59] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. [Online]. Available: https://doi.org/10.1109%2Fcvpr.2015.7298682

[61] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[62] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "Sface: Privacy-friendly and accurate face recognition using synthetic data," 2022. [Online]. Available: https://arxiv.org/abs/2206.10520

[63] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[64] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver io: A general architecture for structured inputs and outputs," 2021. [Online]. Available: https://arxiv.org/abs/2107.14795

[65] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[66] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[67] L. Vaiani, M. La Quatra, L. Cagliero, and P. Garza, "Viper: Video-based perceiver for emotion recognition," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 67–73. [Online]. Available: https://doi.org/10.1145/3551876.3554806

[68] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[69] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[70] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[71] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[72] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*.

[73] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, "The lottery ticket hypothesis for pre-trained bert networks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.