INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multimodal Deception Detection using Automatically Extracted Acoustic, Visual, and Lexical Features

Jiaxuan Zhang[1], *Sarah Ita Levitan*[1], *Julia Hirschberg*[1]

[1]Department of Computer Science, Columbia University, USA
jz2997@columbia.edu, sarahita@cs.columbia.edu, julia@cs.columbia.edu,

Abstract

Deception detection in conversational dialogue has attracted much attention in recent years. Yet existing methods for this rely heavily on human-labeled annotations that are costly and potentially inaccurate. In this work, we present an automated system that utilizes multimodal features for conversational deception detection, without the use of human annotations. We study the predictive power of different modalities and combine them for better performance. We use openSMILE to extract acoustic features after applying noise reduction techniques to the original audio. Facial landmark features are extracted from the visual modality. We experiment with training facial expression detectors and applying Fisher Vectors to encode sequences of facial landmarks with varying length. Linguistic features are extracted from automatic transcriptions of the data. We examine the performance of these methods on the Box of Lies dataset of deception game videos, achieving 73% accuracy using features from all modalities. This result is significantly better than previous results on this corpus which relied on manual annotations, and also better than human performance.

Index Terms: deception, prosody, multimodal data, facial landmarks

1. Introduction

In recent years, deception detection has been increasingly studied by researchers in speech and Natural Language Processing (NLP). Automated methods for deception detection have critical applications for many potential users, including law enforcement, military, intelligence agencies, as well as political and financial organizations. Linguistic cues to deception have been identified in many domains, including fake hotel reviews [1], essays on controversial persuasive essays [2], and interview dialogues [3]. While most of these studies have focused on a single modality, psychologists have also found evidence of visual cues to deception, such as emotional facial expressions [4]. So, combining such modalities has the potential to improve performance of machine learning approaches to detecting deception. Recently, [5] introduced a multimodal Box of Lies dataset of conversational deception from a game played on The Tonight Show Starring Jimmy Fallon®. They collected videos from the show, manually annotated them for language and visual cues, and trained classifiers to detect deception from these manual cues. However, time-consuming manual annotation does not fully explore the potential of automatic deception detection. In this work, we test the use of **fully automatically extracted multimodal features** for truly automated deception detection. The benefits of automatic feature extraction include: reducing the time and cost of annotating an entire video corpus, eliminating the uncertainty of annotation disagreement, increasing the limited utility of manually annotated features, and increasing the potential scale of multimodal methods to very large datasets.

In our work on this publicly available corpus, We leverage methods developed in NLP, speech, and computer vision research to automatically extract linguistic, acoustic and visual features – without any manual annotation or transcription of the data. We evaluate models trained using single modality feature sets and combinations of multimodal features. Our best multimodal classifier achieves an accuracy of 73% using fully automated features, which outperforms [5]'s 65% using human annotations, demonstrating the effectiveness of our approach to automatic multimodal deception detection.

2. Related Work

Identifying differences between deceptive and truthful behaviors is a key challenge in work on deception detection; visual, acoustic, and lexical features have been explored to capture such nuances. Lexical features such as n-grams [6], POS tags [1] and Linguistic Inquiry and Word Count (LIWC) features [2] have been tested in multiple scenarios such as online forum reviews, courtroom trials and human subject experiments to identify cues to deception. Conversational dialogues are particularly useful for studying deception, as they are a common scenario where deception occurs. [3] studied lexical and acoustic cues such as pitch and speaking rate to deception interview dialogues, where two participants were engaged in multi-turn question-answer based conversations. Vision-based features have been found to be useful for deception detection, with [7] and [8] finding facial expressions and body movements to be important features. [9] applied visual deception detection in a conversational setting, finding that participants in an interrogation game exhibited different facial expressions when lying compared to when telling the truth. Multimodal-based deception detection also has had some success in previous studies such as [5] and [6] who found improvements when combining features from different modalities when using manually annotated features. [10] combined improved Dense Trajectory (iDT), MFCC features, and GloVe embeddings to identify deception in real-world trials. While this research has shown that combining features from multiple modalities greatly improved classification performance, the features used in previous approaches have relied on manual data annotations, including human transcription of the speech and meticulous labeling of facial expressions. Inspired by these approaches, we leverage the benefits of multimodal features to build an *automated* system for deception detection in conversational scenarios. However, the only input to our system is a raw video segment; no manual annotation of features is used.

3. The Box of Lies Corpus

We use the Box of Lies corpus [5] for our multimodal approach to deception detection using automatically extracted features

http://dx.doi.org/10.21437/Interspeech.2020-2320

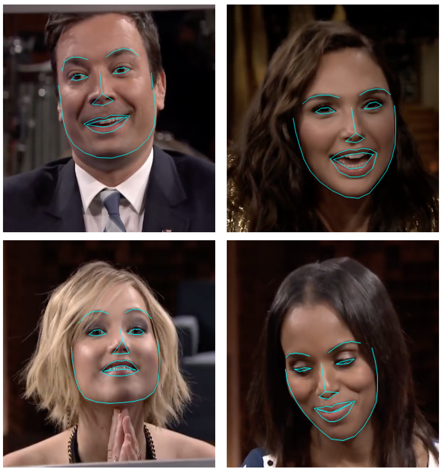

Figure 1: *Facial landmarks detected with different facial expressions. Top left to bottom right: Mouth-Lips Retracted; Head Repeated Tilt; Eyebrows Raising; Smile with Gaze down.*

only. This corpus contains video-recorded conversations between the host of the Tonight Show, Jimmy Fallon, and 28 guests who play a game called "Box of Lies." One player gives a truthful description of an object which they see inside a box or a deceptive description of some fictional object that is not in the box. The other player guesses whether the description is truthful or not. This conversational setting greatly increases the difficulty of predicting truth or lie, since the interaction between participants can be difficult to capture. Also, all players are celebrities, supposedly better actors than normal people, making deception detection even more challenging. There are 68 recorded rounds of the game (39 truth and 29 lie) and a total of 1049 utterances (862 lie and 187 truth). Each round introduces a new box with a new object; the same pair of players can play multiple rounds. The limited size of the corpus at the round level constrains the potential for designing useful features for a prediction model, so most of our experiments were conducted at the utterance level. Each round has a single label of "Truth" or "Lie". However, within a single round with a label of "Lie", it is possible that some utterances were actually truthful. Thus, utterance annotations are more precise. Correct utterance-level labels were provided by the collectors of the corpus. The corpus is unbalanced, with a ratio between truthful and deceptive statements of 1:4.6.

There are key challenges for automatic feature extraction in this corpus: For acoustic features, the recordings have substantial background noise including laughter and applause from the audience as well as background music. For visual features, the angle of the camera often moves, causing changes in the angle of the speakers' faces in different frames. For linguistic features, the two players often speak simultaneously, increasing the difficulty of obtaining high-quality automatic speech recognition (ASR) transcripts to use for linguistic analysis. We address these challenges in our feature extraction approaches below.

4. Methodology

We automatically extracted both verbal and non-verbal features from the Box of Lies Corpus to compare performance on these automated features with performance on manually labeled features in the corpus. Below we describe the methods used to automatically extract features from three modalities: (1) acoustic (2) visual and (3) lexical.

4.1. Acoustic Features

Acoustic-prosodic features have been previously identified as useful for deception detection [11, 12]. However, these studies extracted features from cleanly recorded audio in laboratory settings. Here we extracted features from noisy data. We used the Interspeech 2009 (IS09) ComParE Challenge OpenSMILE baseline feature set [13], a standard benchmark feature set for many computational paralinguistic tasks. Before extracting these features, we first employed multiple noise reduction techniques to clean the data, which contained audience laughter, applause, and music. The noise reduction methods include calculating spectral centroids, MFCCs, and Median filtering, which are described in detail in [14]. We compared the deception detection performance of acoustic features extracted after applying each of the three noise reduction techniques; the results are shown below in Section 5.

4.2. Visual Features

Visual features have been previously explored for deception detection, but they often rely on laborious human annotation. In this work, we explored three approaches to automatically extract visual features: (1) Fisher Vector (FV) encoding (2) Vector of Linearly Aggregated Descriptors (VLAD) encoding and (3) Facial expression detection (FED). First, we used dlib [16] to extract facial landmarks from video frames containing the face of the game describer. To match faces of the same person and exclude the partner from our analysis, we used face embeddings (a 128-D vector representation of the face) to measure the distance between two faces using cosine similarity. To eliminate the effect of faces being located at different positions and different angles, we normalized each distinct face X detected. After identifying the faces of the describer and applying normalization, we explored the following methods to encode visual information as features for deception detection.

FV Encoding. In computer vision, a Fisher Vector (FV) is used to represent an image for classification [15]. A Gaussian Mixture Model (GMM) models the distribution of visual features extracted from an image, and then the FV encodes the gradients of the log-likelihood of the features under the GMM, with respect to the GMM parameters, creating fixed-length vector representations. FVs have been shown to outperform traditional Bag-of-Visual representations for image classification tasks [16].

VLAD Encoding. VLAD is another computer vision approach used to obtain dense vector representations of images [17]. Similar to FV, VLAD encoding also maps features with various lengths to the same latent space. The fundamental difference between FV and VLAD is that VLAD uses K-Means clustering instead of a GMM and only stores first order information.

Facial Expression Detection (FED). In addition to these two unsupervised representations of the images, we also explored a supervised approach for extracting features: facial expression detection. There are 7 categories of facial expressions annotated in the corpus, based on the MUMIN multimodal coding scheme [18]. Each category has multiple behavior values, for example eyebrows can be neutral, rising, frowning, or other. These expressions were carefully annotated in the original corpus by trained annotators. We wanted to leverage those existing gold annotations to train machine learning classifiers to automatically identify expressions on a held-out test set, to explore whether these **automatically predicted** expressions would be more useful for deception detection than our unsupervised methods. We trained individual facial expression detectors using Random Forest classification, using automatically identified facial landmark coordinates as features. For each of the 7 categories, we trained a multi-class facial expression classifier and evaluated the performance using a held out test set. Table 1 shows the performance of each of the 7 multi-class facial expression classifiers.

Facial Expression	Accuracy	F1 Score	# classes
General Face	74%	0.73	5
Eyebrows	75%	0.75	4
Mouth-Lips	73%	0.72	6
Head	67%	0.64	14
Mouth-Openness	82%	0.81	3
Gaze	77%	0.76	6
Eyes	73%	0.72	6

Table 1: *Facial expression classification performance. The detectors are trained using human annotations from training data and evaluated on held out testing data.*

As shown in the table, our classifiers performed well at detecting facial expressions, with results ranging from .64 F1 (Head) to .81 F1 (Mouth-Openness). Head movements were the most difficult to classify and also have the largest number of possible classes (14). These classification results are quite promising, especially given the challenge that the camera angle changes constantly and thus the face of the describer is captured from quite different angles. Also note that the classifier parameters were not tuned for the task to avoid over-fitting. One of the challenges of facial expression detection in this data is the severe data imbalance of data classes. For example, for Mouth-Lips classification, the "Corners down" class has only 305 examples, while other classes such as "Retracted" has ˜3000 examples. In the future, this facial expression detection performance could potentially be improved using sampling or assigning different weights for different classes in the loss function.

In Section 5, we compare the use of FV encodings, VLAD encodings, and automatically identified facial expressions as features for the task of deception detection.

4.3. Linguistic Features

Linguistic features have been successfully used for automatic detection of deceptive speech. However, these are usually extracted from manual transcriptions of speech, which is expensive, time-consuming, and does not scale to large datasets. To automate this process, we generated transcriptions of the dataset using automatic speech recognition (ASR) (Google Cloud Speech-to-Text) and then extract linguistic features from the transcripts. We extracted the following sets of linguistic features from the ASR transcripts: **Unigrams.** Bag-of-words[19] vectors were generated using all the available transcripts. The value at each position in the vector represents the frequency of words in the given utterance; **Psycholinguistic features**. Linguistic Inquiry and Word Count (LIWC) [20] groups words into 80 psychologically-motivated categories and has been successful in many deception detection experiments; **Part of Speech tags.** POS tags for each utterance were generated using NLTK [21], capturing the grammatical and syntactical structure of the utterance as represented as a vector of percentages for each possible tag in the utterance; **Word Embeddings.** Word2Vec [22] was used to extract a vector representation of each word and FV encoding is applied to get the aggregated representation of the utterance to capture semantic relationships between words.

These features were previously explored in the Box of Lies dataset [5] as well as other deception datasets[23] using manual transcriptions. Here, we automated this process using ASR to generate transcriptions and extracting the same features from the transcripts automatically. Our Word-Error-Rate was fairly high at 47.43%, so we were interested to discover how it would serve in our automated multimodal classification tasks as we extracted linguistic features for use in the task.

5. Experiments and Results

Using the automatically extracted feature sets described above, we conducted machine learning deception classification experiments for each modality individually, as well as for the combined modalities. We also compared results from manually generated features with our automated features. For all experiments we used the same experimental settings reported in [5]: a Random Forest classifier with default parameters (scikit-learn implementation [24]) unless otherwise noted. Due to the small size of the dataset, we used five-fold cross-validation to evaluate our models as used in the previous work.

The data is unbalanced across truth and lie classes. 57.35% of game rounds were deceptive, and 82.17% of all utterances spoken were labeled as lies. This is because players were given a choice about whether to lie or tell the truth, and most players chose to lie. To address the class imbalance problem, we applied down-sampling to data labeled as lies to balance the dataset. We also experimented with up-sampling strategies such as Synthetic Minority Over-sampling Technique (SMOTE) [25], but we found that this degraded performance on the Truth class. Since down-sampling was applied on the dataset, a random classifier has a 50% probability of making the correct prediction. We report both accuracy and AUC scores.

Although acoustic-prosodic features have been found to be useful for deception detection in other domains, they have not been previously explored in the Box of Lies dataset, so we cannot compare our performance with previous work on the corpus. However, we compare the performance of models trained using acoustic-prosodic features extracted from both unfiltered noisy audio and from noise-reduced audio (Table 2). As shown in the table, noise reduction was an important pre-processing step for acoustic-prosodic deception classification, achieving the best deception classification result of 63% accuracy using the spectral centroid method; using the unfiltered data we achieved an accuracy of 60%.

			Lie/Truth		
Method	Acc.	AUC	P	R	F1
Spectral Centr.	63%	0.67	.64/.63	.62/.65	.63/.64
MFCC	62%	0.65	.61/.69	.71/.58	.66/.63
Med. Filter	62%	0.63	.63/.62	.60/.64	.62/.63
Orig. (Noisy)	60%	0.63	.61/.59	.57/.63	.59/.61

Table 2: *Deception classification results with acoustic-prosodic features, using original noisy audio and multiple noise reduction methods. Acc. = Accuracy; AUC = Area Under Curve; P = Precision; R = Recall; F1 = F1-score*

Next, we trained deception classifiers using automatically-extracted visual features as described above: facial expressions, FV encoding, and VLAD encoding. Since the dataset also has manual annotations of facial expressions, we compared classification results between our automated features and these manually extracted features (Table 3). As shown here, features generated by our trained facial expression detectors achieved an accuracy of 64%, performing slightly better than our performance on the manually annotated facial expressions, with 62% accuracy. However, the best visual deception classification performance

(67% accuracy) was achieved using FV encoding. Unlike our facial expression detectors, which did rely on human annotations for training, FV encoding and VLAD encoding required no human annotation so this achievement is both impressive and useful for future research, suggesting that automatic classification in this case may be more reliable than human annotation. The performance of FV encoding is slightly better than VLAD encoding, perhaps due to the additional information from the second-order statistics from the FVs.

			Lie/Truth		
Features	Acc.	AUC	P	R	F1
FED	64%	0.72	.65/.63	.61/.67	.63/.65
FV	67%	0.70	.66/.67	.67/.66	.67/.66
VLAD	65%	0.64	.67/.66	.63/.65	.64/.65
MFEA	62%	0.64	.64/.60	.54/.69	.58/.64

Table 3: *Deception classification results using automatically extracted visual features, compared with manual features. FED = Facial Expression Detectors; FV = Fisher Vector Encoding; VLAD = Vector of Linearly Aggregated Descriptors Encoding; MFEA = Manual Facial Expression Annotations*

We also evaluated the performance of each individual lexical feature. Result are shown in Table 4.

Lexical Features	Orig.(Acc./AUC)	ASR(Acc./AUC)
Unigrams	63.3%/0.70	62.0%/0.71
LIWC Features	61.3%/0.68	60.8%/0.68
PoS tags	57.1%/0.65	58.9%/0,62
Word Embeddings	59.7%/0.70	58.5%/0.70
Combined	63.6%/0.73	63.3%/0.73

Table 4: *Deception classification result with lexical features. The "Orig." column shows results using human-labeled transcripts; the "ASR" column shows results using ASR transcripts. The "Combined" feature set uses a combination of all features.*

As shown in Table 4, we found that lexical features extracted from our ASR transcripts performed on par with those extracted from the manual transcripts, despite a relatively high ASR WER. A manual inspection of ASR errors showed that several errors occurred around filler words, which were frequent in this spontaneous speech, suggesting that perfect transcription was not critical for this task; clearly, ASR transcription allowed us to capture important lexical information for deception detection. The best deception detection performance obtained with the ASR transcriptions was 63.3% using a combination of all lexical features; this was only 0.3% lower than the best performance obtained with the gold human transcriptions.

5.1. Multimodal Deception Detection

After experimenting with automatically-extracted features for each modality separately, we trained multimodal classifiers combining the best performing features for each modality. For acoustic features, we used the features extracted from the spectral centroid noise-reduced audio. For visual features, we used FV encoding. For linguistic features, we used a combination of unigrams, POS tags, LIWC, and word embeddings, all extracted from ASR transcripts. We compared our multimodal deception models with the previous manual feature results reported in [5], as shown in Table 5. As shown in the table, combining acoustic, visual, and linguistic modalities achieves the best deception classification performance of 73% accuracy. This is an 8% absolute improvement over the best multimodal performance (65%) achieved using manual features.

In addition to classifying utterances as deceptive or truthful, we also evaluated our proposed method on entire rounds of games. An advantage of evaluating performance at the round level is that it allows us to directly compare our system performance with human performance, since players guess whether their partner is lying or telling the truth after each round. To avoid overfitting on the small amount of round-level data, we trained a classifier on the utterance-level data and then generated the round-level predictions by taking the majority vote on the predictions of all utterances within a single round. Using this utterance-level prediction aggregation approach, our classifier, trained with a combination of automated acoustic, visual, and linguistic features, achieved an accuracy of 72% on entire rounds of the game. This is a 15% absolute improvement over human performance, which was 57%.

			Lie/Truth		
Features	Acc.	AUC	P	R	F1
L+V	71%	0.78	.72/.71	.70/.73	.71/.72
A+V	68%	0.74	.72/.68	.64/.75	.68/.71
L+A	71%	0.74	.71/.68	.71/.73	.71/.71
A+L+V	73%	0.77	.75/.71	.69/.77	.71/.74
Manual L+V	65%	0.68	.64/.66	.67/.63	.66/.65

Table 5: *Multimodal deception classification performance. L = Linguistic; V = Visual; A = Acoustic*

6. Conclusions and Future Research

In this paper, we presented a novel multimodal system for deception detection in conversational dialogue. Importantly, our approach used automatically extracted acoustic-prosodic, visual, and linguistic features that did not require any manual annotation or transcription. We empirically tested the performance of several noise reduction and visual feature extraction approaches to identify the best performing features. In all three individual modalities, we showed that our automatically extracted features perform better than manual annotations. Our final multimodal model combined features extracted from all three modalities, achieving an accuracy of 73%. In addition to improving over previous models trained with gold manual features (65% accuracy) on the Box of Lies utterances, our system also outperformed human performance on the Box of Lies rounds (72% to 57%). This paper contributes to the problem of automatic deception detection in multimodal dialogues, which is an understudied problem. Our work incorporates automatic feature extraction methods developed in computer vision and signal processing and shows how they can be used for multimodal classification of deception.

The Box of Lies dataset provides a useful corpus for deception detection, but it is comprised of a set of videos of a lying game that is played for entertainment. In future work we will evaluate our models on real-world multimodal deception data, such as videos of court testimonies [26] and political speeches, where deception has higher stakes. Unlike human annotations, which are difficult to consistently extract by different annotators in different domains, our automated methods for extracting multimodal features are standardized across domains. In addition, future work can explore more complex classifiers such as recurrent neural networks to model conversational context and time-dependent features to improve automatic deception detection.

7. Acknowledgments

This work was partially funded by AFOSR FA9550-11-1- 0120. We thank Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea for sharing the Box of Lies dataset and annotations.

8. References

[1] 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 - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 309–319. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002512

[2] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[3] 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://www.aclweb.org/anthology/N18-1176

[4] P. Ekman and W. V. Friesen, "Detecting deception from the body or face." *Journal of Personality and Social Psychology*, vol. 29, no. 3, p. 288, 1974.

[5] 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)*, 2019, pp. 1768–1777.

[6] 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: ACM, 2015, pp. 59–66. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820758

[7] T. Fornaciari and M. Poesio, "Automatic deception detection in Italian court cases," *Artificial Intelligence and Law*, vol. 21, pp. 303–340, 2013.

[8] T. O. Meservy, M. L. Jensen, J. Kruse, D. P. Twitchell, J. K. Burgoon, D. N. Metaxas, and J. F. Nunamaker, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems*, vol. 20, pp. 36–43, 2005.

[9] T. K. Sen, M. K. Hasan, Z. Teicher, and M. E. Hoque, "Automated dyadic data recorder (addr) framework and analysis of facial cues in deceptive communication," *ArXiv*, vol. abs/1709.02414, 2017.

[10] Z. Wu, B. Singh, L. S. Davis, and V. S. Subrahmanian, "Deception detection in videos," *CoRR*, vol. abs/1712.04415, 2017. [Online]. Available: http://arxiv.org/abs/1712.04415

[11] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[12] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues," in *Proceedings of Interspeech 2018*, 2018, pp. 416–420. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2443

[13] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[14] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 2016, pp. 670–674.

[15] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 487–493. [Online]. Available: http://dl.acm.org/citation.cfm?id=340534.340715

[16] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3304–3311.

[18] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 273–287, 2007.

[19] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, 12 2010.

[20] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," University of Texas at Austin, Tech. Rep., 2015.

[21] S. Bird, "NLTK: The natural language toolkit," *ArXiv*, vol. cs.CL/0205028, 2002.

[22] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[23] R. Mihalcea and M. Burzo, "Towards multimodal deception detection – step 1: Building a collection of deceptive videos," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 189–192. [Online]. Available: https://doi.org/10.1145/2388676.2388714

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[26] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: Multimodal analysis for deception detection," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 938–943.