





# Multimodal Political Deception Detection

Manvi Kamboj , Christian Hessler , Priyanka Asnani, Kais Riani , and Mohamed Abouelenien , University of Michigan—Dearborn, Dearborn, MI 48128 USA

*Political statements are carefully crafted to garner public support for a particular ideology. These statements are often biased and sometimes misleading. Separating fact from fiction has proven to be a difficult task, generally accomplished by cross-checking political statements against an impartial and trustworthy news source. In this article, we make three contributions. First, we compile a novel multimodal dataset, which consists of 180 videos with accompanying audio recordings and transcripts, featuring 88 politicians categorized by political party. To our knowledge, this is the second multimodal deception detection dataset from real-life data and the first in the political field. Second, we extract features from the linguistic, visual, and acoustic modalities to develop a system capable of discriminating between truthful and deceptive political statements. Finally, we perform an extensive analysis on different multimodal features to identify the behavioral patterns used by politicians when it comes to deception.*

How can the general public distinguish the truth from perceived truths and lies in politics? The need to assess deceit in political statements has influenced the emergence of fact checking as an independent stream in news reporting and, thus, has led to the rise of political fact checking establishments, such as PolitiFact, FactCheck.org, and the Washington Post's fact checker. These fact checking establishments evaluate statements made by politicians in order to determine whether their statements are truthful or deceptive.

Traditional lie detection methods, such as polygraph tests, require the physical cooperation of the individual as well as human judgment. For these reasons, traditional methods cannot aid in the offline analysis of political media. Early work in automatic deception detection relied on data collected in a controlled setting.

In the work by Mihalcea and Strapparava,<sup>1</sup> subjects were instructed to speak about their beliefs in regard to controversial topics, such as death penalty, abortion, and best friend. After which, they were instructed to speak in favor of the opposition to their true beliefs,

thereby providing deceptive statements. While such simulated environments have helped us to better understand the behavior of liars, it is difficult to establish whether these findings would apply to real-life situations. In previous work, the subjects were well aware that they were acting in an artificial setting. This may have altered their natural behavioral responses, particularly their emotional expressions.

The first real-world multimodal high-stake dataset based on court trials was introduced in the work by Pérez-Rosas *et al.*<sup>2</sup> which consisted of 121 videos, and was used for deception detection using textual and gesture modalities. Gogate *et al.*,<sup>3</sup> Carissimi *et al.*,<sup>4</sup> and Wu *et al.*<sup>5</sup> used this dataset and applied multimodal techniques involving acoustic, visual, and lexical analysis for deception detection. The results of these papers showed that multimodal techniques can be successfully used for deception detection for videos recorded in an unimpeded environment.

In this article, we wish to continue our contributions to this field and improve upon our previous work,<sup>2,6-8</sup> by employing multimodal learning-based methodologies. In particular, we make three main contributions. First, we introduce the first multimodal dataset for political deception detection and the second multimodal real-world deception detection dataset, where the other one was based on another domain (court trials).<sup>2</sup> Second, we provide an extensive analysis of the different types of features to

---

1070-986X © 2020 IEEE

Digital Object Identifier 10.1109/MMUL.2020.3048044

Date of publication 29 December 2020; date of current version 29 March 2021.

analyze their capability of detecting political deceit. Third, we build lie detection models using individual and combined modalities.

## RELATED WORK

Traditionally, psychological measures have been used for analyzing human behavior and emotions. However, the fact that deceivers and truth tellers show different cues has motivated researchers to analyze additional features ranging from the physiological, thermal, linguistic, visual, and acoustic modalities.<sup>9–11</sup>

Numerous studies analyzed the relationship between deceptive behavior and linguistic selection. For example, Newman *et al.*<sup>12</sup> examined linguistic indicators of deceit in written stories. In addition, computer vision provides an important tool for covert visual deception detection. Michael *et al.*<sup>13</sup> used the blob analysis to extract information about hands and head movements and augmented it with features extracted from the face. Moreover, psychologists have studied microexpressions to detect lies. Since they appear for such a small fraction of time, it is extremely difficult to hide these types of expressions.

More recently, researchers introduced multimodal approaches for creating more informed deceit detection systems. This aims to avoid the uncertainty related to the use of single modalities and presents the benefit of enriching the dataset with information from different sources.<sup>14</sup>

## DATASET

Our goal is to build a multimodal collection of high-stake occurrences of real-life political deception. There is no existing dataset to detect political deceit in the context of a learning-based multimodal system, to the best of our knowledge. A lexical dataset related to fake news detection was introduced by Wang.<sup>15</sup> Their work was based on statements classified by the fact-checking website PolitiFact.com. The dataset is only useful for lexical analysis, as it lacks any multimodal input. In order to bridge this gap, we propose a novel dataset derived from multimodal data extracted from a number of political statements from various politicians.

PolitiFact.com provides a label for each statement, indicating its degree of truthfulness. PolitiFact collects its facts from various news sources to promote political transparency. In doing so, PolitiFact has defined an extensive and well-established set of criteria to evaluate the truthfulness of political statements. A recent study addressed this issue by performing a comprehensive text analysis on approximately 10,000 PolitiFact articles to identify biased treatment of Democrats versus

Republicans.<sup>16</sup> The authors found no obvious differences in the language used to describe members of each party that would indicate bias or differential treatment.

## Data Collection

Our dataset features Donald Trump, Hillary Clinton, Bernie Sanders, Barack Obama, Marco Rubio, Paul Ryan, and Ted Cruz among many other politicians. We focused on collecting only political statements that include all the three modalities: audio, lexical, and visual. Videos were downloaded from different public resources. We imposed several constraints during the data collection process to avoid introducing extraneous information or “noisy data” into our results. We also limited the number of videos to a maximum of six per politician in order to avoid any bias.

The first task was to search for videos of political statement that are classified by PolitiFact. The following task was to locate and extract the segment of the video clip containing the statement. Additionally, the subject’s face needed to be clearly visible in the video, especially while the statement was being made. Furthermore, the video quality had to be of a certain standard to extract meaningful visual features. In general, the video had to have a minimum resolution of approximately 480p, had no excessive camera motion, and featured only one subject of interest.

The audio feature extraction process required the subject’s voice to be clearly audible without any music or noise in the background. There are many classified statements in PolitiFact that are based on political advertisements. We had to exclude those videos as they included loud music playing in the background. In gathering the transcripts, we had to search for the source of the statement and, then, manually reconstruct the complete statement. This was due to the fact that PolitiFact often paraphrases statements instead of providing word for word transcriptions.

Given the challenges and our set goal, the data collection process was very demanding and arduous involving several iterations of web data mining, cleaning, and analysis. The final dataset for our experiments consists of 180 videos with an average length of 18.65 s of 88 different politicians, with two videos overlapping with other videos, including one linguistically mismatched video. The distribution between parties is 87 Democratic videos and 93 Republican videos. The partywise distribution of data is shown in Figure 1.

We collected data for the following six classes of deception and truthfulness as rated by PolitiFact: Pants on Fire, False, Mostly False, Partly True, Mostly True, and True. Our data sheet contains details for the video id, source, date when the statement was made,

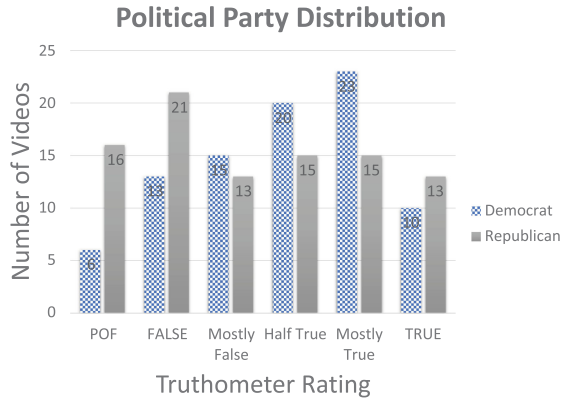


FIGURE 1. PolitiFact truthometer rating.

flag to indicate if the statement was made in a prepared (in a speech etc.), or spontaneous (in a debate, interview, etc.) setting, the political party, transcript, gender, and duration of the video. The dataset includes 146 male subject videos and 34 female subjects, which is due to the imbalanced distribution of gender in the political field. We plan to make this dataset publicly available.

## METHODOLOGY

### Linguistic Modality

Previous work in the literature have effectively demonstrated the relationship between certain linguistic features and deceptive rhetoric. Therefore, we extracted similar linguistic features from the manual transcripts of each video. We ensured that the written transcripts match the video recording. In particular, the following techniques were used.

*Linguistic Inquiry and Word Count (LIWC):* To extract features related to the psychological state of a subject when stating the truthful and deceitful statements, we selected LIWC. LIWC is a transparent textual analysis tool for creating word counts for psychologically meaningful groupings and has been used widely in textual deceit detection. It extracts the cognitive, structural, and emotional components present in text. We used LIWC 2015, which has a dictionary containing almost 6,400 words, select emoticons and word stems, and produces 90 features for each statement received as an input. The 90 features are grouped into 7 broad subcategories, such as summary language variables (which relate to emotional tone, analytical thinking, clout, and authenticity), standard linguistic dimensions (verbs, nouns, etc.), general descriptors, psychological concepts-related words, personal concerns (home, work), informal language

markers, and punctuation markers. To generate our feature set using LIWC, we extracted the frequency of words of every category for each video transcript.

*Part-Of-Speech (POS) Tagging:* We used the NLTKs POS tagger to generate the POS tags for the linguistic features. The approach uses a greedy averaged perceptron tagger, trained on the Wall Street Journal corpus. For our linguistic features, the extracted POS features are encoded using frequency distribution for each POS tag in the dataset.

*Semantic Features:* We used Global Vectors for Word Representation (GloVe)<sup>17</sup> for creating a global vector for each video transcript. GloVe is an unsupervised learning algorithm developed by Stanford for generating word embeddings. The embeddings are created using statistics derived from global word–word co-occurrence in a corpus using log bilinear regression model using both global matrix factorization and local context window models. We used the Wikipedia 2014+ Gigaword5 pretrained corpus with word embedding vector of size 100. The transcripts from the videos were first lemmatized, and then, corresponding word embedding vectors were created using the GloVe corpus.

*Unigrams:* We used bag-of-words representation of transcripts to extract unigram counts, which were used as linguistic features. We started by removing the punctuations from all the text documents and building a vocabulary consisting of all words appearing in the transcripts. The features are encoded as a word–frequency pair, where each word is associated with a value corresponding to the frequency of the unigram inside the transcripts. The final feature set consisted of 2,029 unique words. Note that we also attempted to use tf-idf (i.e., term frequency–inverse document frequency) representation of transcripts to understand the importance of each word with respect to the whole corpus and eliminate words with low scores. However, the evaluation did not show any further improvement.

*Sentiment Polarity:* We used Valence Aware Dictionary and Sentiment Reasoner (VADER) to perform sentiment analysis on the transcripts to determine whether a political statement is positive, negative, or neutral in order to analyze whether sentiment correlates with deception. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiment expressed in social media. It uses a sentiment lexicon—A list of lexical features (e.g., words) that are labeled based on their semantic orientation as positive or negative. We obtained four polarity indices for each of the transcripts “positive,” “negative,” “neutral,” and “compound.” The “positive,”

“negative,” and “neutral” scores are ratios for proportions of text that fall in each category while the “compound” score is a metric that computes the sum of all the lexicon ratings, which have been normalized between  $-1$  (most extreme negative) and  $+1$  (most extreme positive).

### Acoustic Modality

We used OpenSMILE (i.e., open-source media interpretation by large feature-space extraction) to extract acoustic features. In previous research in detecting deception in spoken dialog, researchers compared OpenSMILE and Praat for classification. Levitan<sup>18</sup> identified the top 20 features from OpenSMILE. Hence, we created an acoustic feature set using these top 20 features.

In addition, in order to distinguish between deceptive and truthful speech, we examined the INTERSPEECH 2009 and 2013 Emotion Challenge feature sets (IS09 & IS13) used for emotion recognition. Previous work in the literature have used emotional scores to predict deception.<sup>19</sup> Accordingly, we decided to use the acoustic emotion features to predict deception.

The IS09 feature set<sup>20</sup> contains 32 descriptors that are divided into 16 delta regression coefficients and 16 low-level descriptors (LLD). These include 12 mel-frequency cepstral coefficients (MFCCs 1-12), pitch frequency (F0), zero-crossing-rate, root-mean-square frame energy, and Noise-to-Harmonics ration. Twelve functionals were applied: arithmetic mean, standard deviation, skewness, kurtosis, maximum and minimum value, range, relative position of max. and min. value, linear regression slope, offset, and quadratic error. We extracted 384 features from the audio using the IS09 feature set.

The INTERSPEECH 2013 (IS13) ComParE Challenge feature set contains 6,373 features from the computation of various functionals over LLD contours. Among the paralinguistic tasks, this feature set was used for deception detection.

The final feature set is a combination of the IS13 and IS09 feature sets. This combination yields a set of 6,727 features after eliminating redundant features.

### Visual Modality

OpenFace is used to extract facial behavior features from the videos. It uses constrained local neural field (CLNF) for facial landmark detection. CLNF represents an improvement over constrained local model (CLM), which struggled to perform in poor lighting condition, the presence of blockage, among other challenges. The CLNF includes a local neural field (LNF) patch

expert, which learns about both the adjacent and long-distance pixels by gaining information about the similarity and the long-distance sparsity constraints. This provides local variation of each landmark’s appearance. The second main component for facial landmark detection is point distribution model, which captures variation in the shape of facial landmarks. When processing videos, OpenFace initializes the CLNF model based on facial landmarks detected in previous frames. This provides the detection of 68 facial landmarks.

To get the head pose, the three-dimensional detected facial landmarks are projected on the image, using an orthographic camera projection. For detecting eye gaze, eye-region landmarks are detected first, and then, the pupil location is calculated based on the intersection of a ray passed through the pupil and the eye ball sphere.

To generate facial appearance features, histogram of oriented gradients (HOG) descriptors are extracted from the aligned face in the form of a high dimensional vector (4,464-dimensional vector). After which, PCA is applied in order to reduce the dimensionality. The lower-dimensional HOG and facial features from CLNF are used for action units (AU) prediction. In particular, the presence and intensity of some AU are detected. The final vector representing each video was computed by calculating the mean, maximum, and standard deviation of the frame-based AU, gaze and head pose features. Furthermore, we detected different emotions from the videos using combinations of AU. Each frame of the video is marked with the presence of four basic emotions based on the occurrence of the AU combination required for that emotion. The most frequent occurrence of an emotion is marked as the emotion for the whole video. If no emotion is recognized in a given frame, we marked it as neutral. Hence, each video has one of the four basic emotions (happy, sad, surprise, anger) or neutral as an additional feature at the end of this process.

### Classification

After extracting features from individual modalities, we tested their performance independently and combined. The multimodal fusion was achieved by integrating the features collected from two or three multimodal streams to create a single feature vector, which was then utilized to classify the video. A decision tree classifier was used for classification, as recommended in previous lie detection research.<sup>7</sup> We experimented other classifiers for the lie detection problem; however, decision tree was consistently providing improved results. In addition, deep

**TABLE 1.** Top 10 LIWC classes for Republicans and Democrats (Using Binary 1 classification scheme).

Order	Republican		Democrat	
	False	True	False	True
1	Dic	Dic	Dic	Dic
2	Clout	Clout	Analytic	Clout
3	Analytic	Analytic	Clout	Analytic
4	WC	Function	Tone	WC
5	WPS	Tone	Function	WPS
6	Function	WC	WC	Function
7	Authentic	Authentic	WPS	Authentic
8	Tone	WPS	Authentic	Tone
9	Sixltr	Sixltr	verb	Sixltr
10	Verb	Verb	Sixltr	Relativ

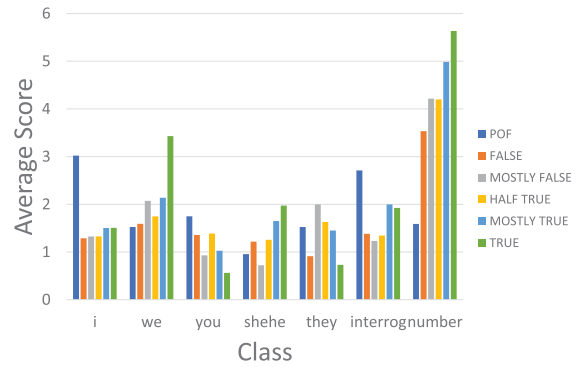
learning was not suitable based on the size of this type of datasets. A leave-one-subject-out cross validation scheme was used, where all instances that belong to the same politician were held in reserve for testing, whereas all the instances belonging to the remaining politicians were used for training in each fold. This scheme avoids any possible bias during training by avoiding the presence of recordings from the same subject in both training and testing. We report the average overall accuracy as well as the recall of each of the deceptive and truthful classes. Recall refers to the average accuracy per class, which is determined by dividing the number of correctly classified instances per class by the total number of instances belonging to that class.

### EXPERIMENTAL DISCUSSION

Given our novel dataset of 180 instances for linguistic, acoustic, and visual modalities, we started by evaluating classification results for deceit detection from individual modalities and, then, assessed their combinations. The dataset is represented in six different categories as provided by PolitiFact. Hence, we converted them into three different binary classification schemes as follows.

- › Binary Combinations
  - › Binary 1: POF + FALSE (as Deceptive) and MOSTLY TRUE + TRUE (as Truthful).
  - › Binary 2: POF + FALSE + MOSTLY FALSE (as Deceptive) and MOSTLY TRUE + TRUE (as Truthful).
  - › Binary 3: POF + FALSE + MOSTLY FALSE (as Deceptive) and HALF TRUE + MOSTLY TRUE + TRUE (as Truthful).

LIWC: Linguistic and other grammar



**FIGURE 2.** LIWC class distribution.

### Linguistic

Table 1 shows the top 10 LIWC classes distribution for Republicans and Democrats. One notable observation from Table 1 is the higher value of emotional tone by Democrats when stating a false statement compared to a true statement. A higher value of tone is associated with a more positive style, implying that Democrats express more positive statements when lying compared to telling the truth. An opposite trend is observed in statements made by Republicans. In general, it is observed that the same LIWC classes are used for deceptive and truthful statements made by politicians of both parties. This can be contributed to the fact that most statements are prepared in advance, which makes it more difficult to distinguish lies in political statements.

Some other interesting indications are shown in Figure 2, which depicts the comparison of interesting LIWC categories. It clearly depicts the use of the self-reference singular pronoun “I” when stating a lie (POF rating) compared to the higher use of plural pronoun “we” when stating a truthful statement (TRUE rating). Another noticeable insight is the higher use of interrogatives (how, when, what) when the statement is marked as POF. The figure also shows a higher use of numbers in truthful statements compared to deceitful statements. As numbers are inclined to be associated with stating facts, we can see their association with more truthful scenarios.

Table 2 shows the classification results using different linguistic sets. The three binary schemes achieve accuracy figures in the 60 s. In particular, GLOVE, POS, and polarity, as well as their combinations exhibit improved performance compared to other sets. Unigrams and LIWC exhibit deteriorated performance, which again can be attributed to making

**TABLE 2.** Linguistic classification results Comb. 1: Glove + POS Comb. 2: Glove + POS + Polarity Comb. 3: Polarity Scores + Unigrams + POS + Glove + LIWC (Accuracy and recall are reported as percentage values).

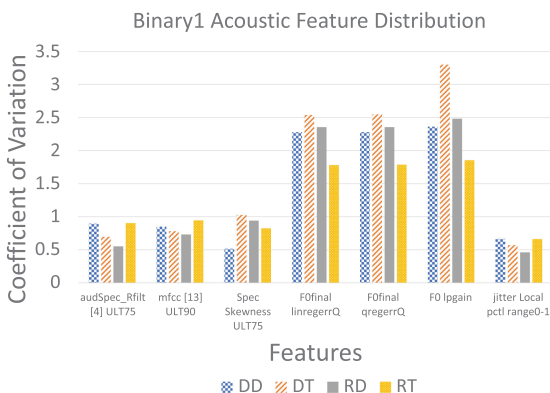
Set	Binary 1			Binary 2			Binary 3		
	Acc.	Recall		Acc.	Recall		Acc.	Recall	
		D	T		D	T		D	T
GLOVE	<b>61</b>	61	61	52	45	58	46	57	31
LIWC	43	41	44	54	56	53	52	57	44
Polarity	58	52	64	53	46	58	57	58	56
POS	55	59	51	53	48	58	<b>61</b>	65	54
Unigrams	53	57	49	53	49	56	46	52	38
Comb. 1	60	57	62	52	50	54	59	62	56
Comb. 2	60	55	64	50	44	55	56	64	44
Comb. 3	47	52	43	<b>66</b>	63	68	47	56	34

Highest accuracy is highlighted in bold.

prepared statements. It can be noticed that the Binary 1 combinations (linguistic feature sets) achieve the best performance in five out of eight cases, followed by Binary 2. This indicates that the performance is potentially related to the combination. Binary 1 makes the best discrimination between the PolitiFact labels, as it avoids using "Half True" and "Mostly False." Binary 2 avoids "Half true." On the other hand, despite the fact that Binary 3 is learning from more data by using all labels, it achieves lower performance, as it provides poorer discrimination between truths and lies.

**Acoustic**

Figure 3 shows some of the acoustic features for "Binary 1" that exhibited interesting opposite patterns between Republicans and Democrats when it comes to deception, such as MFCC and F0 using the coefficient of variation calculated as the standard deviation divided by the mean. In particular higher coefficients of F0 features exhibited a higher association with truthful statements by Democrats (DT) compared to their deceptive statements (DD). On the contrary,



**FIGURE 3.** Acoustic features.

**TABLE 3.** Acoustic features classification results (Accuracy and recall are reported as percentage values).

Set	Binary 1			Binary 2			Binary 3		
	Acc.	Recall		Acc.	Recall		Acc.	Recall	
		D	T		D	T		D	T
20	51	54	48	41	46	34	44	36	51
IS09	<b>53</b>	61	47	58	63	52	54	60	49
IS13	46	46	45	<b>63</b>	70	53	51	54	49
IS09+IS13	43	39	47	62	71	48	<b>56</b>	54	58

Highest accuracy is highlighted in bold.

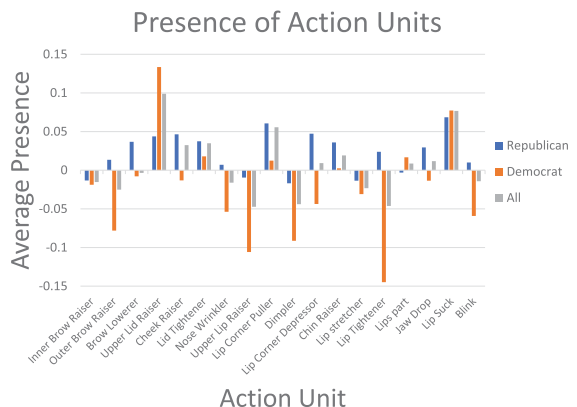
higher values had higher association with deceptive Republicans (RD) compared to truthful ones (RT).

Table 3 lists the accuracy and recall results for different binary combinations schemes for the acoustic features. The results indicate that different feature sets yield different results for different binary schemes. IS13 and IS09 show improved performance compared to the 20 features identified by Levitan. <sup>18</sup> In particular, IS13 outperforms all other sets using the "Binary 2" combination.

**Visual**

Figure 4 provides more insight on how deceptive and truthful statements associate with the AU features extracted from the videos. To find an indication of deceptive versus truthful expressions, the graph bars are calculated by subtracting the deceptive AU average feature values from the truthful ones. Therefore, a positive result specifies an association between an AU and truthfulness, and a negative result indicates an association between an AU and deception.

The resulting figure provides interesting observations. The Cheek Raiser and Lip Corner Puller are mostly associated with truthful statements. Similarly, Upper lid raiser and Lip suck are strongly associated with truthful statements for all the subjects irrespective of their party. It is interesting how Republicans' Lip



**FIGURE 4.** Action units.

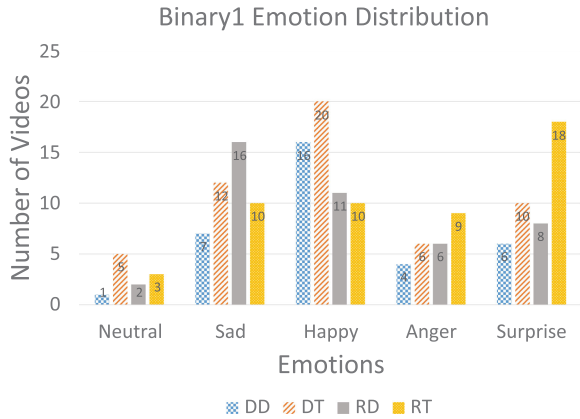


FIGURE 5. Visual emotion features.

tightener is associated with truthful statements, whereas for Democrats, it is associated with lying. Similar behavior is seen for Lip corner depressor between the parties, with Republicans showing its association with truth while Democrats showing it for deception.

In order to gain more insight on the difference between truthful and deceptive statements, Figure 5 shows the distribution of the five aforementioned emotions. Interestingly, the figure shows there is a higher association between sadness and deceptive Republicans. On the contrary, it shows higher association between sadness and truthful Democrats. A similar trend is observed for happiness. The other emotions indicate similar patterns between the two parties with different intensities.

Table 4 shows that visual features: AU and Gaze, and specifically some of their combinations, may be strong indicators for deception detection. In particular, combining multiple visual features improves the classification accuracy. The feature set consisting of AU, Gaze and Pose achieve the highest accuracy of 68%.

### Multimodal

Different combinations of linguistic, acoustic, and visual features are experimented and compared for performance. Table 5 lists the classification results using various multimodal feature set combinations. One notable observation is that the feature sets consisting of two modalities frequently outperform those consisting of three modalities if the acoustic modality is excluded. The highest accuracy is attained by performing the "Binary 1" classification using only lexical and visual features. This particular combination of the visual eye gaze and emotion with the linguistic POS reaches close to 70% accuracy, which is the highest among all individual and combined modalities.

TABLE 4. Visual AU classification results Comb. 1: AU + Gaze Comb. 2: AU + Gaze + Pose Comb. 3: Gaze + Emotion (Accuracy and recall are reported as percentage values).

Set	Binary 1			Binary 2			Binary 3		
	Acc.	Recall		Acc.	Recall		Acc.	Recall	
		D	T		D	T		D	T
AU	62	46	75	55	82	18	57	39	72
Emotion	41	34	48	49	80	07	56	42	69
Gaze	56	59	61	54	64	49	<b>61</b>	63	60
Pose	49	48	49	50	58	39	53	56	51
Comb. 1	62	59	64	64	70	56	<b>61</b>	57	65
Comb. 2	<b>68</b>	64	72	59	70	43	52	50	53
Comb. 3	62	57	66	<b>65</b>	67	62	<b>61</b>	67	56

Highest accuracy is highlighted in bold.

TABLE 5. Fusion (All Features): [visual All; Acoustic IS09+IS13; Lexical All] Fusion1 (3 Modalities): [visual AUs, Gaze, Pose; Acoustic; Lexical Glove] Fusion2 (3 Modalities): [visual Gaze, Emotion; Acoustic IS13; Lexical Glove] Fusion3 (3 Modalities): [visual Gaze, Emotion; Acoustic IS13; Lexical All] Fusion1 (2 Modalities): [visual Gaze, Emotion, Pos] Fusion2 (2 Modalities): [visual Gaze, Emotion; Lexical Glove] Fusion3 (2 Modalities): [visual Gaze, Emotion; Lexical Polarity].

Set	Binary 1			Binary 2			Binary 3		
	Acc.	Recall		Acc.	Recall		Acc.	Recall	
		D	T		D	T		D	T
Fusion-ALL	48	48	48	63	67	57	57	54	59
Fusion1-3M	56	55	58	50	56	37	52	49	55
Fusion2-3M	55	53	57	61	66	53	56	52	58
Fusion3-3M	46	44	48	61	65	55	<b>58</b>	55	59
Fusion1-2M	<b>69</b>	63	75	61	66	53	52	49	55
Fusion2-2M	58	56	60	<b>65</b>	70	58	53	50	56
Fusion3-2M	66	63	69	63	66	57	57	53	60

Highest accuracy is highlighted in bold.

One rational explanation for this is that the acoustic features may have less discriminative power, as it relates to deception detection, compared to the lexical and visual features. It is also likely that the sheer number of approximately 6,000 acoustic features increased the variance in the feature set. This large number of features may have introduced a bias toward acoustic properties while undermining the lexical and/or visual properties. It can also be noticed that the same trend of an improved performance of Binary 1 and Binary 2, compared to Binary 3, is observed. For instance, Binary 1 reaches 69% accuracy, whereas Binary 3 does not exceed 58%.

In order to perform additional analysis, we expanded the binary classification into a three-class classification using Deceptive (POF+False+Mostly False), Half True, and True (Mostly True+True) with decision tree in Table 6(A) and into a six-class ordered logistic regression problem without merging in Table 6(B). The highest output regression probabilities for the instances are converted into predictions,

**TABLE 6.** Results Using (A) trinary classification using deceptive (POF+False+Mostly False), Half True, and True (Mostly True+True) and (B) ordered logistic regression using all six classes without merging (*Accuracy and recall are reported as percentage values*).

Set	Overall Accuracy	Deceptive	Half True	True
Baseline		47	19	34
Fusion-ALL	36	45	21	31
Fusion1-3M	36	51	13	30
Fusion2-3M	32	45	11	27
Fusion3-3M	32	44	10	29
Fusion1-2M	42	<b>55</b>	21	37
Fusion2-2M	35	48	7	35
Fusion3-2M	<b>43</b>	52	<b>30</b>	<b>39</b>

(A)

Set	Acc.	POF	False	M. False	H. True	M. True	True
Baseline		12	19	16	19	21	13
Fusion-ALL	19	<b>27</b>	09	21	23	21	<b>13</b>
Fusion1-3M	18	05	12	21	46	13	00
Fusion2-3M	17	05	06	32	49	03	00
Fusion3-3M	19	05	06	36	54	05	00
Fusion1-2M	22	14	15	25	46	<b>24</b>	00
Fusion2-2M	22	00	09	36	<b>60</b>	13	00
Fusion3-2M	<b>23</b>	00	09	<b>57</b>	<b>60</b>	05	00

(B)

Highest accuracy and recall are highlighted in bold.

which are used to calculate the accuracy and recall in the table. This expansion, however, results in a lower number of instances per class. As expected the results for both sections of the table indicate that the reduction in the number of instances per class deteriorates the overall performance. However, in many cases the recall per class exceeds the baseline in the tables, which is specified based on random guessing. The results also agree with the pattern noticed earlier in Table 5 that the fusion of the visual and linguistic modalities has better capability of modeling lies and truths, compared to the inclusion of the acoustic modality, which again can be attributed to the reasons mentioned earlier. Another interesting observation is the improved modeling of the “Half True” instances, in particular, compared to other classes using regression, which indicates that regression is able to separate instances with features that cannot be exclusively specified as truths or lies. This observation, however, is not conclusive due to the low number of instances per class in this specific case and needs future exploration.

## CONCLUSION

Political deception detection is becoming increasingly important due to ethical concerns and to raise public awareness. We introduced a novel multimodal dataset for political deception detection. To the best of our knowledge, there is no other multimodal dataset available for this specific task. The statements made in the videos were not restricted to a specific topic and

several constraints were placed during the data collection process to avoid any bias.

We provided an analysis of the linguistic, visual, and acoustic features and observed interesting behavioral differences between deceptive and truthful politicians as well as between Republicans and Democrats. For example, some facial AUs and emotions had higher association with deceptive statements and specific parties. Moreover, our linguistic analysis uncovered an inverse relationship, in which emotional tone is associated with truthful statements made by Republicans, and on the contrary, with deceitful statements made by Democrats.

Furthermore, we extracted multiple types of feature sets to construct classifiers in order to automatically detect political lies. Different truthometer combinations were experimented to create multiple binary classification schemes. The classification results indicated the potential of automatically detecting political lies using multimodal features, reaching approximately 70% accuracy with the integration of the visual and linguistic modalities in particular.

The results indicate that developing an automated multimodal political deception detection system can aid in determining the truthfulness of the statements and promises made by politicians and can potentially make them think twice before stating a lie or manipulating facts. In future work, we plan to use additional feature selection approaches in order to find the optimal set of features from each modality and construct more advanced models to classify political statements. We also plan to use transfer learning to explore the feasibility of using domain-specific lie detection models.

## REFERENCES

1. R. Mihalcea and C. Strapparava, “The lie detector: Explorations in the automatic recognition of deceptive language,” in *Proc. ACL-IJCNLP Conf. Short Papers*, 2009, pp. 309–312.
2. V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” in *Proc. Int. Conf. Multimodal Interact.*, 2015, pp. 59–66.
3. M. Gogate, A. Adeel, and A. Hussain, “Deep learning driven multimodal fusion for automated deception detection,” in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2017, pp. 1–6.
4. N. Carissimi, C. Beyan, and V. Murino, “A multi-view learning approach to deception detection,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 599–606.
5. Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, “Deception detection in videos,” in *Proc. AAAI Conf. Artif. Intell.*, 2018.



6. V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo, "Verbal and nonverbal clues for real-life deception detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2336–2346.
7. M. Abouelenien, V. Prez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 5, pp. 1042–1055, May 2017.
8. M. Abouelenien, M. Burzo, V. Prez-Rosas, R. Mihalcea, H. Sun, and B. Zhao, "Gender differences in multimodal contact-free deception detection," *IEEE Multimedia*, vol. 26, no. 3, pp. 19–30, Jul.–Sep. 2019.
9. S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics: Short Papers—Volume 2*, 2012, pp. 171–175.
10. Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis, "Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1280–1289, May 2013.
11. B. A. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 6, pp. 1015–1023, Jun. 2014.
12. M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality Social Psychol. Bull.*, vol. 29, no. 5, pp. 665–675, 2003.
13. N. Michael, M. Dilsizian, D. Metaxas, and J. K. Burgoon, "Motion profiles for deception detection using visual cues," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 462–475.
14. M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: Multimodal analysis for deception detection," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops*, 2016, pp. 938–943.
15. Y. Wang, "Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, vol. 2, pp. 422–426.
16. Dallas Card, L. H. Lin, and N. A. Smith, "PolitiFact language audit," Mar. 2018. [Online]. Available: <https://homes.cs.washington.edu/~nasmith/papers/card+lin+smith.tr18.pdf>
17. J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
18. S. I. Levitan, "Deception in spoken dialogue: Classification and individual differences," Ph.D. dissertation, New York, NY, USA: Columbia Univ., 2019.
19. S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. W. Schuller, "Is deception emotional? An emotion-driven predictive approach," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2011–2015.
20. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

**MANVI KAMBOJ** is currently a Product Development Engineer with Ford. Her research interests include multimodal interaction, machine learning, and natural language processing. She received the master's degree from the University of Michigan—Dearborn, Dearborn, MI, USA. Contact her at [mkamboj@umich.edu](mailto:mkamboj@umich.edu).

**CHRISTIAN HESSLER** is currently a Senior Embedded Software Engineer with General Motors. His research interests include machine learning, computer vision, and natural language processing. He received the master's degree from the University of Michigan—Dearborn, Dearborn, MI, USA. Contact him at [cahessle@umich.edu](mailto:cahessle@umich.edu).

**PRIYANKA ASNANI** is currently an Economic Analyst with the Federal Reserve Bank of Dallas, Dallas, TX, USA. Her research interests include machine learning, natural language processing, and econometrics. She received the master's degree from the University of Michigan—Dearborn, Dearborn, MI, USA. Contact her at [pasnani@umich.edu](mailto:pasnani@umich.edu).

**KAIS RIANI'S** research interests include deep learning, computer vision, natural language processing, and artificial intelligence. He received the engineering degree from "Ecole Polytechnique de Tunisie," Marsa, Tunisia. He is currently working toward the Ph.D. degree with the University of Michigan—Dearborn, Dearborn, MI, USA. Contact him at [kriani@umich.edu](mailto:kriani@umich.edu).

**MOHAMED ABUELENIEN** is currently an Assistant Professor with the Department of Computer and Information Science, University of Michigan—Dearborn, Dearborn, MI, USA. He was a Postdoctoral Research Fellow with Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI, from 2014 to 2017. His research interests include multimodal deception detection, multimodal sensing of thermal discomfort and drivers' alertness levels, emotion and stress analysis, machine learning, ensemble learning, image processing, face and action recognition, and natural language processing. He received the Ph.D. degree in computer science and engineering from the University of North Texas, Denton, TX, USA, in 2013. He is the corresponding author of this article. Contact him at [zmohamed@umich.edu](mailto:zmohamed@umich.edu).