electronics

MDPI

Article

Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach

Muhammad Shehram Shah Syed *, Elena Pirogova and Margaret Lech

School of Engineering, RMIT University, Melbourne, VIC 3000, Australia; elena.pirogova@rmit.edu.au (E.P.); margaret.lech@rmit.edu.au (M.L.)

* Correspondence: muhammad.shehram.shah.syed@rmit.edu.au

Abstract: This paper explores the automatic prediction of public trust in politicians through the use of speech, text, and visual modalities. It evaluates the effectiveness of each modality individually, and it investigates fusion approaches for integrating information from each modality for prediction using a multimodal setting. A database was created consisting of speech recordings, twitter messages, and images representing fifteen American politicians, and labeling was carried out per a publicly available ranking system. The data were distributed into three trust categories, i.e., the low-trust category, mid-trust category, and high-trust category. First, unimodal prediction using each of the three modalities individually was performed using the database; then, using the outputs of the unimodal predictions, a multimodal prediction was later performed. Unimodal prediction was performed by training three independent logistic regression (LR) classifiers, one each for speech, text, and images. The prediction vectors from the individual modalities were then concatenated before being used to train a multimodal decision-making LR classifier. We report that the best performing modality was speech, which achieved a classification accuracy of 92.81%, followed by the images, achieving an accuracy of 77.96%, whereas the best performing model for text-modality achieved a 72.26% accuracy. With the multimodal approach, the highest classification accuracy of 97.53% was obtained when all three modalities were used for trust prediction. Meanwhile, in a bimodal setup, the best performing combination was that combining the speech and image visual modalities by achieving an accuracy of 95.07%, followed by the speech and text combination, showing an accuracy of 94.40%, whereas the text and images visual modal combination resulted in an accuracy of 83.20%.

Keywords: trust classification; social signal processing; speech acoustics; computer vision; multimodal fusion

check for updates

Citation: Syed, M.S.S.; Pirogova, E.; Lech, M. Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach. *Electronics* **2021**, *10*, 1259. https://doi.org/10.3390/electronics10111259

Academic Editor: Antonio Orlandi

Received: 18 April 2021
Accepted: 21 May 2021
Published: 25 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Trust is an important social characteristic, guiding interactions between people in society. In general terms, it is the willingness to subject oneself to the actions of other individuals. The popularity of politicians, for example, is a measure of public trust in specific individuals. In continuation of our previous work [1,2], we performed experiments for the automatic prediction of public trust using audio data. In these works, we found that there was a significant statistical difference in the speech characteristics of politicians deemed highly trustworthy and others that were trusted less. Moreover, we also observed important gender-based differences [1]. A multilayer perceptron was used to classify speech data into low-trust and high-trust politicians with an average accuracy of 81% being achieved. From our findings, it was concluded that speech characteristics could be used for the prediction of public trust in politicians. In the subsequent study [2], we proposed a multimodal framework for predicting trust using speech and text modalities, and we therein evaluated trust prediction using the two modalities individually, as well as their combination. The results showed that using more than one modality for trust prediction led to significant improvements in classification performance. The current study is an extension of our investigation into the multimodal prediction of trust, where we proposed

an enhanced multimodal prediction framework and extended experiments by including a mid-trust category, as well as adding a third modality (image).

Here, we further investigated objective indicators for the prediction of public trust manifested through social signals, and we extended our dataset by including the image modality that complements the two modalities already available, i.e., speech and text. We employed advanced machine learning (ML) techniques used in social signal processing (SSP) [3,4] for speech analysis by representing speech data as standard acoustic feature-sets in the OpenSMILE toolkit [5]. Text features were represented using natural language processing (NLP) techniques, such as the bag of words/term frequency—inverse document frequency (BoW/TF-IDF) [6], document to vector (doc2vec) [7], as well as variations in the BERT [8] and distilbert [9] models. Image features were represented using computer vision models such as ResNet50 [10], VGG16 [11], and Xception [12]. These formats/models represent state-of-the-art techniques for each of the respective modalities.

In this study, a novel multimodal framework was suggested for classification and its performance was evaluated through thorough experimentation (Figure 1). To validate experiments conducted for the prediction of trust, a baseline was first established using the modalities of speech, text, or images individually. The results from the proposed multimodal fusion of the three modalities were then compared to the achieved baseline results. The multimodal fusion was a two-stage process. The first stage involved predicting the trust via each modality individually (Figure 1, Stage 1). We stored the generated class probability values and labels generated in this step for use in the next stage. In the second stage, the confidence values or labels for all individual modalities in the first stage were combined to form a single fusion vector. This fused vector was passed as an input for the training of a separate fusion classifier, which was the ultimate decision maker for the final prediction of trust (Figure 1, Stage 2).

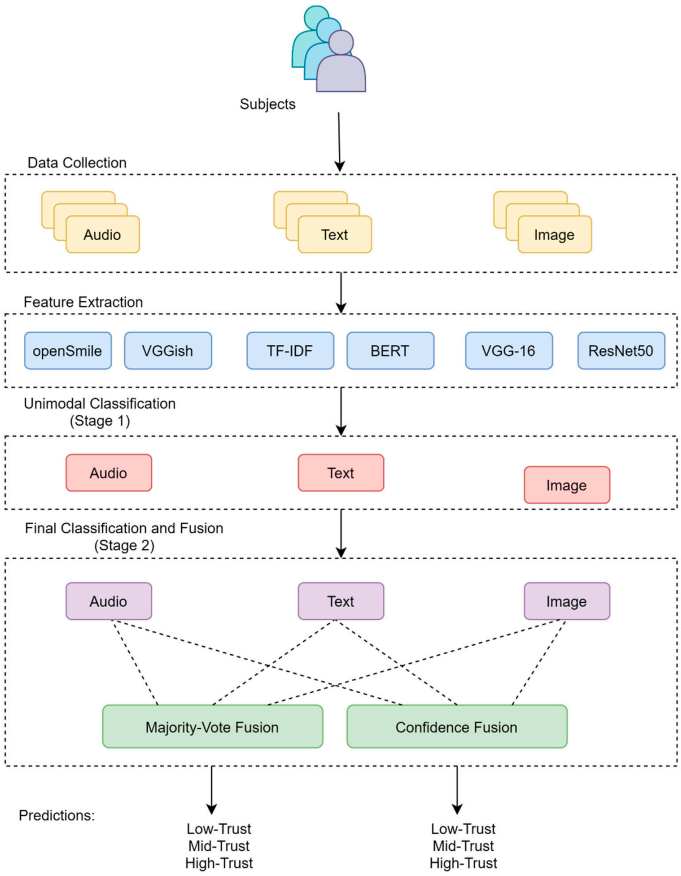

Figure 1. Experimental framework: Stage 1—Unimodal Classification; Stage 2—Final Classification and Fusion (Multimodal).

The rest of the paper is structured as follows. Section 2 provides a review of the literature for tasks utilizing speech, text, or images for social signal processing applications. Section 3 discusses the database preparation and feature representation process of speech, text, and image databases. Section 4 elucidates on the methodology. The experiments and their results are discussed in Section 5, with a conclusion being provided in Section 6.

2. Related Works

Advancements in computing technologies have led to the development of sophisticated technologies for the analysis of human behavior. Aspects of human behavior, commonly referred to as social signals, can be expressed in the form of spoken or written words, body language, or facial expressions [13]. Our behavior is a manifestation of the conscious and subconscious aspects of cognitive processes. Amongst various artefacts of human conduct, speech, text, and visual attributes have been widely used in various research areas derived from a larger study domain, i.e., behavioral analysis.

The most prominent forms of human communication are text and speech, and understandably, both these forms have been used in research extensively to determine traits related to human behavior. Our mental state is directly related to the manner of the production of our speech. Research studies have indicated that the same part of the human brain controls social interactions and speech [14]; therefore, insights into social behavior can be provided by analyzing speech data.

Various applications pertaining to the areas of engineering, computer science, and psychology have found speech acoustics to be useful. Examples of such applications include emotion detection [15–20], analysis of human behavior and personality [21,22], mood detection [23], and even being used for the prediction of the chances of employment [24]. Specifically, in the areas of mental wellbeing and health, speech, text, and visual information have been used for screening different types of disorders, including developmental, cognitive, behavioral, emotional, and psychological. Moreover, these modalities have also been used as assistive technologies to screen for depression [21,23,25,26], schizophrenia [27–29], Alzheimer's disease [18,20,30,31], bipolar disorder [32,33], and autism spectrum disorders (ASD) [16,34–36].

Most studies in the field of engineering, focused on investigating trust, have been based on speech analysis [37], where it has often been contrasted with deception [15,38,39]. Some studies have also investigated textual/lexical features [40,41], as well as facial expressions [42]. The authors in [43] converted speech recordings into RGB images of spectrograms; then, using a computer vision model for trust prediction, they predicted trust from images, but those images represented speech. These studies suggest that speech, text, and visual attributes may provide insights into trustworthiness. Whereas these studies utilized a unimodal prediction, we proposed a multimodal system for trust prediction.

It has been reported that fusing different modalities has the potential to provide more information compared to a single modality [32–34]. In fact, some research works have explored complementary and cross-modality information infusion, as reported in [44,45]. The authors in [46] provided a review of methods for multimodal data fusion. To this end, our study explored the use of two approaches for data fusion, involving multiple modalities (speech, text, and images). The prediction resulting from multiple combinations of these three modalities was compared to results produced individually from each modality.

3. Database Generation and Feature Computation

A diverse range of social signals have, for a long time, been investigated to find clues and gain insights into human behavior. Examples of these include speech, text, facial expressions, and gestures, which have been widely used by researchers for a variety of behavioral analysis tasks. Even though the prediction of trust using speech, text, and images has been extensively studied by psychologists, this field is in its nascent state within the field of engineering. Furthermore, to the best of our knowledge, no publicly available dataset exists for use in trust prediction. Therefore, we have created our own dataset for

the task at hand. We also made use of publicly available results of the online ranking of public trust in USA politicians [47]. Publicly available speech, Tweets (text), and pictures of politicians taken from the ranking list and trust labels were generated using their ranking scores. The collected data were recorded no longer than 12 months prior to the publishing of the ranking scores in January 2018. The final dataset consisted of speech, text, and image data of fifteen well-known politicians. The individuals were divided into three groups: the first one included politicians perceived to have high-trust, the second one included politicians in the mid-trust category, and the third category included low-trust politicians. Each class consisted of data from five individuals, two of which were males, and the other three were females. In addition, only those politicians who had a minimum of 100 votes were considered in this study. The three trust categories were formed from the individuals based on the ratio of positive votes cast vs. the total number of votes cast.

3.1. Speech Modality

The speech database consisted of 30 audio files for each of the fifteen politicians included in the study. These audio files had an average length of 12 min and were extracted from the audiovisual recordings of these politicians. Acoustic speech features were calculated using the OpenSMILE [5] toolkit; this toolkit has become the industry standard for speech acoustic research. The speech modality was represented as acoustic parameters using five commonly used parameter sets for paralinguistic research. The names of these feature sets are:

1. eGeMAPS (Geneva Minimalistic Acoustic Parameter Set) [48];
2. ComParE (COMputational PARalinguistics challengE) [49];
3. IS09 Emotion (Interspeech 2009) [50];
4. IS10 Paralinguistics (Interspeech 2010) [51];
5. Prosody [5].

3.2. Text Modality

The text database consisted of 30,325 tweets (High Trust: 10,755, Mid Trust: 9804, Low Trust: 9766). It was created by processing *Tweets* (from *Twitter)* of the same politicians included in the speech database. A number of data processing-included tasks were performed to make it suitable for classification experiments. Major tasks included removing all formatting elements, punctuation marks, and special characters.

Text data were represented using a variety of NLP techniques. These included features represented as Bag of Words (BoW) [6] and document to vector (doc2vec) [7,52], which are some of the most basic but widely used text representation techniques. The other sets of features were derived using pretrained text embeddings of several variations of Bidirectional Encoder Representations from Transformers (BERT) and a distilled version of BERT [8,9] pretrained networks. The names of these features/models are as follows:

1. Bag of Words/Term Frequency- Inverse Document Frequency (BoW/TF-IDF);
2. Document to Vector (Doc2Vec);
3. BERT base uncased;
4. BERT large uncased;
5. DistilBERT uncased;
6. bert-base-nli-stsb-mean-tokens;
7. bert-base-nli-mean-tokens;
8. distilbert-base-nli-stsb-mean-tokens.

3.3. Image Modality

A visual database was created by downloading images of subjects from Google using an image scrapping tool. A total of 2196 images (High Trust: 752, Mid Trust: 741, and Low Trust: 703) were downloaded and processed. Classification experiments were performed using three pretrained image processing models. The names of the models, which represent the current state of the art of computer vision tasks, are as follows:

1. Resnet50;
2. Xception;
3. VGG16.

4. Methodology

The proposed framework performs "end-to-end" tasks right from data preparation and feature extraction up to data assembling and the final prediction of trust. As mentioned earlier, the multimodal framework predicts trust in two stages (Figure 1). In the first stage, trust is predicted on the basis of each modality individually. This stage is a pre-requisite to the final prediction of trust, which is based on using multiple modalities, and we refer to it as the second stage.

Stage 1: Moving on to a "step-by-step" description, raw data for each modality are fed into the system. In the first step, we represent each modality in a suitable format, and we have used a variety of state-of-the-art models for each modality. In the next step, we perform a unimodal prediction of trust using each modality individually. Here, we save the classification labels, as well as the confidence values, which we shall later use for information fusion. In the final step, we fuse the input values (confidence and labels) from Stage 1 to create a new feature vector, which will be used to train the final decision-making classifier.

Stage 2: The fusion of the modalities is performed using two techniques, confidence-based fusion and label-based fusion. In confidence-based fusion, the final prediction of trust category is determined on the basis of the average of confidence scores of the unimodal classifier for predicting each class, whereas, for label-based fusion, we simply perform a majority vote of the prediction outcomes (labels) of each modality.

The block diagram (Figure 1) shows a general framework of the data processing and classification steps applied in our experiments.

Classification Performance Measure

The classification performance of the framework has been measured using the accuracy a defined as:

$$a = \frac{t_p + t_n}{t_p + t_n + f_n + f_n} \tag{1}$$

where t_p and t_n are the numbers of true-positive and true-negative classification outcomes, respectively. Similarly, f_p and f_n denote the numbers of false-positive and false-negative classification outcomes, respectively. The final accuracy is computed as an average of fivefold cross-validation.

5. Experiments and Results

The experiments were divided into three categories: (i) unimodal prediction of trust, (ii) bimodal prediction of trust, and (iii) trimodal prediction of trust. Unimodal prediction refers to the use of any single modality from speech, text, and images to achieve the prediction of trust; bimodal prediction refers to the use of any two modalities to determine the prediction of trust, whereas trimodal prediction refers to the use of all three modalities to determine the prediction of trust. We discuss the specific aspects of experimental design and the obtained results for each experiment separately, as follows.

5.1. Unimodal Prediction of Trust

The proposed multimodal fusion framework uses confidence values and prediction labels generated by single modality classifiers working with either of the three modalities. Therefore, the first stage of our experiments implemented the unimodal classification. Three separate logistic regression (LR) classifiers were trained and tested, one each for speech, text, and images.

5.1.1. Trust Prediction from Speech

The speech dataset was divided into the training (80%) and testing (20%) subsets, and a, LR classifier was trained to differentiate between individuals with low-trust, mid-trust, and high-trust, using acoustic speech features computed previously (Section 3.1). The class probability vectors and labels, generated during the testing stage, were saved for future processing in the multimodal classification, as described in Section 4. Table 1 shows speech-based classification results given as an accuracy for five standard acoustic feature-sets. The results are presented as an average over five runs through fivefold cross-validation. We report that the highest accuracy of 92.81% on the test partition was achieved with the *ComParE* feature set, closely followed by the *IS10* feature-set, resulting in an accuracy of 92.32%. It is interesting to note that although *IS10* had a much smaller number of features than those of *ComParE*, they had a similar performance. The choice of feature-set can be made on the basis of the experimental goal—whether one aims to achieve the highest accuracy or one aims for a relatively lightweight but high-performing feature-set. In our study, we opted for the best performing model in terms of the classification accuracy, which is the *ComParE* feature-set.

Table 1. Classification results showing three-class trust prediction using speech acoustics.

Feature Set	No. of Features	Accuracy Validation (%)	Accuracy Test (%)
ComParE	6373	93.28	92.81
IS10	1582	92.88	92.32
IS09	384	82.99	82.72
eGeMAPS	88	80.51	80.39
Prosody	25	58.65	58.86

5.1.2. Trust Prediction from Text

Trust prediction from text was performed using the same LR setup, which was used for the speech data features computed using each of the embeddings and representation formats, as mentioned in Section 3.2. The dataset of text messages was divided into the training (80%) and testing (20%) subsets, and the classifier was trained to distinguish between the individuals with low-trust, mid-trust, and high-trust, using text features. The results are shown in Table 2. As can be seen, the highest accuracy of 72.31% on the test partition was achieved with the *BERT_large_uncased* model, followed closely by the *distilbert_uncased* model, resulting in an accuracy of 72.26%. Comparing the top four best performing NLP formats, one can see that their performance was largely similar. However, there were major underlying differences amongst these models. The foremost is that the first three techniques use pretrained word embeddings to represent data, whereas the *BoW/TF-IDF* generates feature vectors from scratch. Comparing the three pretrained NLP models, one notes that they differ considerably in architecture. For matters of simplicity, we only compared the number of parameters in each model. *BERT_large_uncased* comprises 336 million parameters, *distilbert_uncased* has 66 million parameters, whereas *BERT_base_uncased* has 110 million parameters. It is interesting to note here that *BoW/TF-IDF,* where we trained the model from scratch, also provided results comparable to the pretrained models.

Table 2. Prediction of trust through text.

Model	Accuracy Validation (%)	Accuracy Test (%)
BERT_large_uncased	72.29	72.31
distilbert_uncased	72.36	72.26
BERT_base_uncased	71.33	71.48
BoW/TF-IDF	71.31	70.99
bert-base-nli-stsb-mean-tokens	67.65	67.38
bert-base-nli-mean-tokens	67.64	67.17
distilbert-base-nli-stsb-mean-tokens	67.27	66.86
distilbert-base-nli-mean-tokens	67.01	66.58
doc2vec	64.77	64.78

5.1.3. Trust Prediction from Images

Trust prediction from images was performed using the same LR setup, which was used for the speech and text data analysis. In this case, the images were fed to the networks listed in Section 3.3. The dataset of images was divided into the training (80%) and testing (20%) subsets, and the classifier was trained to distinguish between the individuals with low-trust, mid-trust, and high-trust. The results are shown in Table 3. *Resnet50* provided the highest classification accuracy of 77.96%. It considerably outperformed the other two computer vision models used for this task.

Table 3. Prediction of trust through images.

Model	Accuracy Validation (%)	Accuracy Test (%)
ResNet50	77.83	77.96
Xception	71.62	71.11
VGG16	69.82	69.55

5.1.4. Comparison of Individual Modalities

A comparison between the three modalities showed that speech on its own led to a higher overall performance compared to text and images. The best performing features for each modality are presented in Table 4.

Table 4. Best performing features for each modality.

Modality	Feature Name	Accuracy Validation (%)	Accuracy Test (%)
Speech	ComParE	93.28	92.81
Text	distilbert_uncased	72.36	72.26
Image	ResNet50	77.83	77.96

5.2. Multimodal Prediction of Trust

As shown above, we demonstrated that any modality from either speech, text, or images can be used on its own to predict trust. The next question is whether we can improve trust prediction by combining additional modality/modalities. To answer this question, we created a new set of multimodal features by concatenating the probability vectors and labels generated by the unimodal classifiers to train a decision-making classifier for classification based on multiple modalities, as shown in Figure 1. This third (bimodal) or fourth (trimodal) classifier acted as the decision maker to determine the final trust label. It is important to mention here that only the best performing features of each modality have been used for the multimodal prediction of trust.

Assuming that vectors $\mathbf{p}_{ai} = \{p_{a1i}, p_{a2i}\}$ for $i - 1, \ldots, N$ (where N is the number of data samples) represent two-class probability vectors generated by the LR trained on speech

data, and $\mathbf{p}_{bi} = \{p_{b1i}, p_{b2i}\}$ are the two-class probability vectors generated by the LR trained on text data, multimodal feature vectors $\mathbf{p}_{sti}$ were generated as

$$\mathbf{p}_{abi} = \{p_{a1i}, p_{a2i}, p_{b1i}, p_{b2i}\} \text{ for } i - 1, \ldots, N \quad (2)$$

Similarly, if we were to add image modality represented as $\mathbf{p}_{ci} = [p_{c1i}, p_{c2i}\}$ to the above system, the resultant vector for multimodal prediction using three modalities would be generated as

$$\mathbf{p}_{abci} = \{p_{a1i}, p_{a2i}, p_{b1i}, p_{b2i}, p_{c1i}, p_{c2i}\} \text{ for } i - 1, \ldots, N \quad (3)$$

Having the knowledge of the ground truth labels for each sample i, we were able to train a third LR classifier to perform trust recognition based on the multimodal feature vectors, $\mathbf{p}_{abi}$, representing any two modalities from speech, text, and images or $\mathbf{p}_{abc}$ when all three modalities are used for trust prediction. Based on this, we investigated multiple combinations of modalities to achieve multimodal prediction. The combinations of the three modalities for multimodal prediction are as follows:

- Speech + Text
- Text + Images
- Images + Speech
- Speech + Text + Images

As the number of examples for each modality was different, we fed an equal number of samples to the prediction framework, so vector fusion could be formed. As discussed earlier, the multimodal prediction of trust was achieved in two stages: the first stage predicting trust using a singular modality, and the second stage predicting trust through multiple modalities by vector concatenation (as shown in Equations (2) and (3)). As the size of dataset changed, we again performed unimodal prediction of trust with the truncated dataset to maintain consistency of the classification process. The results of stage 1 are shown in Table 5.

Table 5. Stage 1 results of multimodal prediction of trust.

Modality	Feature Name	Accuracy Validation (%)	Accuracy Test (%)
Speech	ComParE	93.39	92.73
Text	distilbert_uncased	76.83	77.93
Image	ResNet50	69.67	68.67

When combining the classification measures of multiple modalities, there are several approaches to achieve multimodal prediction. One such approach is confidence level fusion where we concatenate the confidence or probabilities of each class for every modality to create a new feature vector, as shown in Equations (2) and (3). Another approach is to achieve final multimodal prediction through voting of the labels of individual modalities, with the final prediction being determined by the label receiving the most votes from individual modalities. However, this technique cannot be applied when only two modalities are at play, as a majority vote cannot be achieved, due to an even number of modalities. The selection of technique usually depends on the number of modalities being used. Therefore, for multimodal prediction involving two modalities, the final prediction is achieved using confidence level fusion, whereas for multimodal prediction using three modalities, we achieve the final prediction using confidence level fusion, as well as the majority voting of labels.

5.2.1. Multimodal Prediction of Trust Using Two Modalities (Bimodal Prediction)

Bimodal prediction refers to the use of any two modalities to achieve a multimodal prediction of trust. The results of using different combinations of the three modalities in a bimodal setup are given in Table 6. The best performing combination was "image and speech," which provided the highest accuracy of 95.07% on the test partition. This

was closely followed by "speech and text," resulting in an accuracy of 94.40% on the test partition. These results outperformed the highest accuracy of 92.73% achieved with speech data only in unimodal prediction (stage 1 results, see Table 5). As can be observed from Table 6, speech was one of the two best performing modalities in a bimodal setup, thus suggesting that it is a very strong modality for trust prediction.

Table 6. Results of multimodal prediction of trust using combinations of two modalities.

Modality	Accuracy Validation (%)	Accuracy Test (%)
Speech + Text	94.33	94.40
Text + Images	82.00	83.20
Images + Speech	95.28	95.07

In order to further investigate the combination of different representation formats of each modality, we performed additional experiments using the top 2 and top 3 performing models of each modality to determine how different combinations of features result in trust prediction outcomes. The results are given in Tables 7 and 8, respectively. It can be observed from the two tables that, as in the previous experiment, speech was present in the best performing results. Furthermore, it can also be noted that experiments including the text modality produced the worst results, which follows from its poor performance on an individual level. Another thing to note is that, using the top 3 models (Table 8) for each modality produced a slight reduction in performance compared to the top 2 model experiment (Table 7). This can be attributed to the phenomenon of diminishing returns where adding more models does not result in improved results.

Table 7. Results of bimodal prediction of trust using top 2 performing models.

Modality	Accuracy Validation (%)	Accuracy Test (%)
Speech + Text	96.28	96.07
Text + Images	85.17	85.13
Images + Speech	97.61	97.53

Table 8. Results of bimodal prediction of trust using top 3 performing models.

Modality	Accuracy Validation (%)	Accuracy Test (%)
Speech + Text	95.67	95.53
Text + Images	84.02	83.88
Images + Speech	96.42	96.58

5.2.2. Multimodal Prediction of Trust Using Three Modalities (Trimodal Prediction)

Trimodal prediction refers to the use of all three modalities for the prediction of trust. The results are shown in Table 9. As discussed earlier, we used two techniques, i.e., confidence-based fusion and label-based fusion, to achieve the final prediction of trust. It can be seen that the label fusion (majority voting) technique results in an accuracy of 95.33% on the test partition, considerably outperforming the accuracy of 91.87% achieved with confidence fusion.

Table 9. Results of multimodal prediction of trust using combinations of three modalities.

Fusion Technique	Accuracy Validation (%)	Accuracy Test (%)
Confidence Fusion	92.11	91.87
Majority Voting	95.78	95.33

When comparing the results shown in Tables 4 and 6, Tables 7–9, we can conclude that combining multiple modalities improved the classification accuracy. This is in line with

the findings of our previous studies. To further explore the trimodal prediction of trust, experiments were conducted using the top 2 and top 3 performing models for all three modalities. The results are given in Tables 10 and 11, respectively. It can be observed that using the top 2 models provided a marginally better result compared to using the top 3 models, as was observed in Section 5.2.1.

Table 10. Results of trimodal prediction of trust using top 2 performing models.

Fusion Technique	Accuracy Validation (%)	Accuracy Test (%)
Confidence Fusion	97.06	97.53
Majority Voting	94.89	95.27

Table 11. Results of trimodal prediction of trust using top 3 performing models.

Fusion Technique	Accuracy Validation (%)	Accuracy Test (%)
Confidence Fusion	95.73	95.91
Majority Voting	93.89	94.13

6. Conclusions

This study provides a comprehensive investigation into the suitability of three modalities of social signals, i.e., speech, text, and images, for the prediction of trust. It evaluated the effectiveness of each modality individually but also explored multiple combinations of these modalities to achieve trust prediction. It proposed a multimodal framework for trust prediction that performs "end-to-end" tasks. The results revealed that by using multiple modalities, we can improve the classification accuracy compared to a single modality. The multimodal classification approaches presented in this paper enabled us to compare and contrast between different combinations of the three modalities (and their respective representation formats) to determine the most suitable approach for trust prediction.

The proposed framework for the multimodal prediction of trust was utilized to demonstrate its usefulness to effectively predict public trust in politicians by using three different data modalities (speech, text, and images). Through the experiments, we were able to ascertain that it was possible to achieve a relatively high prediction accuracy for all three modalities i.e., 92.81% for speech, followed by 77.93% for text, and 68.67% for images. However, when these modalities were combined, the accuracy could be increased by up to 97%. This was achieved by using the top 2 models for all three modalities combined. Moreover, when comparing the results of different combinations of modalities in multimodal experiments, the highest accuracy was achieved when speech was one of the modalities used. This suggests that speech is a very robust modality for the task at hand. This can be attributed to the fact that speech is one of the oldest modalities to be used for behavioral analysis and social signal processing tasks. Subsequent advancements in speech analysis techniques have resulted in the development of many mature technologies. However, an assessment of trustworthiness that is based on more than one modality would be the natural choice, as multiple modalities provide more robustness in the prediction framework and make the assessment more reliable. This is also true for people in a real-world environment; when we are able to see and hear a person and read their texts, we are able to form a more comprehensive opinion about the person as compared to assessing them only on the basis of their speech, written text or even just an image.

One of the critical limitations of the methodology employed here is that it applies only to a closed set of politicians. In future studies, we intend to investigate the utility of the proposed framework by extending it for a more generalized dataset.

Author Contributions: Conceptualization, M.S.S.S., E.P. and M.L.; methodology, M.S.S.S., E.P. and M.L.; software, M.S.S.S.; validation, M.S.S.S., E.P. and M.L.; formal analysis, M.S.S.S.; investigation, M.S.S.S.; resources, M.S.S.S. and M.L.; data curation, M.S.S.S.; writing—original draft preparation, M.S.S.S.; writing—review and editing, M.S.S.S., E.P. and M.L.; visualization, M.S.S.S.; supervision, E.P. and M.L.; project administration, E.P. and M.L.; funding acquisition, M.S.S.S., E.P. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The final dataset is planned to be made available for researchers upon completion of research project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Syed, M.S.S.; Stolar, M.; Pirogova, E.; Lech, M. Speech Acoustic Features Characterising Individuals with High and Low Public Trust. In Proceedings of the 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS), Gold Coast, QLD, Australia, 27 February 2019; pp. 1–9.
2. Syed, M.S.S.; Pirogova, E.; Lech, M. Multimodal Prediction of Public Trust in Politicians from Speech and Text. In Proceedings of the 2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS), Adelaide, Australia, 14–16 December 2020; pp. 1–6.
3. Vinciarelli, A. *Introduction: Social Signal Processing*; Cambridge University Press: Cambridge, UK, 2017.
4. Vinciarelli, A.; Pantic, M.; Bourlard, H. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.* **2009**, *27*, 1743–1759. [CrossRef]
5. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Seoul, Korea, 25 October 2010; pp. 1459–1462.
6. Zhang, Y.; Jin, R.; Zhou, Z.-H. Understanding bag-of-worDS model: A statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [CrossRef]
7. Le, Q.V.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 22–24 June 2014; pp. 1188–1196.
8. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
12. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
13. Vinciarelli, A.; Pantic, M.; Bourlard, H.; Pentland, A. Social signals, their function, and automatic analysis: A survey. In Proceedings of the 10th International Conference on Multimodal Interfaces, Crete, Greece, 20–22 October 2008; pp. 61–68.
14. Yap, T.F.; Epps, J.; Ambikairajah, E.; Choi, E.H.C. Voice source under cognitive load: Effects and classification. *Speech Commun.* **2015**, *72*, 74–95. [CrossRef]
15. Herms, R. Prediction of Deception and Sincerity from Speech Using Automatic Phone Recognition-Based Features. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 2036–2040.
16. Holbrook, S.; Israelsen, M. Speech Prosody Interventions for Persons with Autism Spectrum Disorders: A Systematic Review. *Am. J. Speech Lang. Pathol.* **2020**, 1–17.
17. Stolar, M.N.; Lech, M.; Bolia, R.S.; Skinner, M. Real time speech emotion recognition using RGB image classification and transfer learning. In Proceedings of the Signal Processing and Communication Systems (ICSPCS), 2017 11th International Conference on, Gold Coast, Australia, 13–15 December 2017; pp. 1–8.
18. Syed, M.S.S.; Syed, Z.S.; Lech, M.; Pirogova, E. Automated Screening for Alzheimer's Dementia through Spontaneous Speech. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020.
19. Theodoros, D.G.; Hill, A.J.; Russell, T.G. Clinical and quality of life outcomes of speech treatment for Parkinson's disease delivered to the home via telerehabilitation: A noninferiority randomized controlled trial. *Am. J. Speech Lang. Pathol.* **2016**, *25*, 214–232. [CrossRef] [PubMed]
20. Weiner, J.; Herff, C.; Schultz, T. Speech-Based Detection of Alzheimer's Disease in Conversational German. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 1938–1942.
21. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]

22. Schirmer, A.; Chiu, M.H.; Lo, C.; Feng, Y.; Penney, T.B. Angry, old, male—and trustworthy? How expressive and person voice characteristics shape listener trust. *PLoS ONE* **2020**, *14*, e0210555.
23. Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; Pantic, M. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 21 October 2016; pp. 3–10.
24. Nguyen, L.S.; Frauendorfer, D.; Mast, M.S.; Gatica-Perez, D. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans. Multimed.* **2014**, *16*, 1018–1031. [CrossRef]
25. Girard, J.M.; Cohn, J.F.; Mahoor, M.H.; Mavadati, S.; Rosenwald, D.P. Social risk and depression: Evidence from manual and automatic facial expression analysis. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8.
26. Williamson, J.R.; Quatieri, T.F.; Helfer, B.S.; Ciccarelli, G.; Mehta, D.D. Vocal and facial biomarkers of depression based on motor incoordination and timing. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 3–7 November 2014; pp. 65–72.
27. Nagels, A.; Kircher, T.; Grosvald, M.; Steines, M.; Straube, B. Evidence for gesture-speech mismatch detection impairments in schizophrenia. *Psychiatry Res.* **2019**, *273*, 15–21. [CrossRef] [PubMed]
28. Tron, T.; Peled, A.; Grinsphoon, A.; Weinshall, D. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In Proceedings of the International Symposium on Pervasive Computing Paradigms for Mental Health, Milan, Italy, 24–25 September 2015; pp. 72–81.
29. Tron, T.; Peled, A.; Grinsphoon, A.; Weinshall, D. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In Proceedings of the 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 24–27 February 2016; pp. 220–223.
30. Fraser, K.C.; Rudzicz, F.; Hirst, G. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, 27 June 2016; pp. 1–11.
31. Haider, F.; De La Fuente, S.; Luz, S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 272–281. [CrossRef]
32. Ringeval, F.; Schuller, B.; Valstar, M.; Cowie, R.; Pantic, M. Summary for AVEC 2018: Bipolar disorder and cross-cultural affect recognition. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 25 October 2018; pp. 2111–2112.
33. Syed, Z.S.; Sidorov, K.; Marshall, D. Automated screening for bipolar disorder from audio/visual modalities. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, Seoul, Korea, 22 October 2018; pp. 39–45.
34. Guha, T.; Yang, Z.; Ramakrishna, A.; Grossman, R.B.; Darren, H.; Lee, S.; Narayanan, S.S. On quantifying facial expression-related atypicality of children with autism spectrum disorder. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing/Sponsored by the Institute of Electrical and Electronics Engineers Signal Processing Society, ICASSP (Conference), Queensland, Australia, 19–24 April 2014; p. 803.
35. Oller, D.K.; Niyogi, P.; Gray, S.; Richards, J.; Gilkerson, J.; Xu, D.; Yapanel, U.; Warren, S. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13354–13359. [CrossRef]
36. Samad, M.D.; Diawara, N.; Bobzien, J.L.; Harrington, J.W.; Witherow, M.A.; Iftekharuddin, K.M. A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 353–361. [CrossRef]
37. Belin, P.; Boehme, B.; McAleer, P. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE* **2017**, *12*, e0185651. [CrossRef]
38. Burgoon, J.K.; Stoner, G.; Bonito, J.A.; Dunbar, N.E. Trust and deception in mediated communication. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003; p. 11.
39. Levitan, S.I.; Maredia, A.; Hirschberg, J. Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 416–420.
40. Kopev, D.; Ali, A.; Koychev, I.; Nakov, P. Detecting Deception in Political Debates Using Acoustic and Textual Features. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019.
41. Mendels, G.; Levitan, S.I.; Lee, K.-Z.; Hirschberg, J. Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1472–1476.
42. DeBruine, L.M. Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proc. R. Soc. B Biol. Sci.* **2005**, *272*, 919–922. [CrossRef] [PubMed]
43. Sandoval, C.; Panganiban, A.R.; Stolar, M.; Bolia, R.; Lech, M. Prediction of Inter-Personal Trust and Team Familiarity from Speech: A Double Transfer Learning Approach. *IEEE Access* **2020**, *8*, 225437–225447.
44. Sui, J.; Adali, T.; Yu, Q.; Chen, J.; Calhoun, V.D. A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods* **2012**, *204*, 68–81. [CrossRef]

45. Wagner, J.; Andre, E.; Lingenfelser, F.; Kim, J. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans. Affect. Comput.* **2011**, *2*, 206–218. [CrossRef]
46. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **2020**, *32*, 829–864. [CrossRef] [PubMed]
47. Rothschild, M. The Most Trustworthy Politicians. Available online: //www.ranker.com/list/trustworthy-politicians/mike-rothschild (accessed on 1 March 2018).
48. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
49. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E. The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of the Interspeech 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013.
50. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
51. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S.S. The Interspeech 2010 paralinguistic challenge. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
52. Reddy, D.M.; Reddy, D.N.V.S.; Reddy, D.N.V.S. Twitter Sentiment Analysis using Distributed Word and Sentence Representation. *arXiv* **2019**, arXiv:1904.12580.