Transfer Learning between Concepts for Human Behavior Modeling: An Application to Sincerity and Deception Prediction

Qinyi Luo+, Rahul Gupta+, Shrikanth Narayanan*

+Department of Electronic Engineering, Tsinghua University, Beijing, China
*Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

Abstract
Transfer learning (TL) involves leveraging information from sources outside the domain at hand for enhancing model performances. Popular TL methods either directly use the data or adapt the models learned on out-of-domain resources and incorporate them within in-domain models. TL methods have shown promise in several applications such as text classification, cross-domain language classification and emotion recognition. In this paper, we propose TL methods to computational human behavioral trait modeling. Many behavioral traits are abstract constructs (e.g., sincerity of an individual), and are often conceptually related to other constructs (e.g., level of deception) making TL methods an attractive option for their modeling. We consider the problem of automatically predicting human sincerity and deception from behavioral data while leveraging transfer of knowledge from each other. We compare our methods against baseline models trained only on in-domain data. Our best models achieve an Unweighted Average Recall (UAR) of 72.02% in classifying deception (baseline: 69.64%). Similarly, applied methods achieve Spearman’s/Pearson’s correlation values of 49.37%/48.52% between true and predicted sincerity scores (baseline: 46.51%/41.58%), indicating the success and the potential of TL for such human behavior tasks.

Index Terms: Transfer Learning (TL), sincerity prediction, deception prediction

1. Introduction
Transfer Learning (TL) focuses on leveraging knowledge from one domain and applying it to other related domains [1]. In particular, domains with limited resources (e.g. the ones with an absence of comprehensive annotated datasets) can gain from TL approaches as it is possible to learn models on a related problem and adapt the models to domain of interest. Human behavior modeling [5, 6] is one such domain where obtaining reliable (human) annotations on large datasets is both expensive and prohibitive in time. In this paper, we consider the problems of modeling deceptive intent (deceptive vs not-deceptive) and rating perceived sincerity; applying TL methods in assisting one problem using resources from the other. We hypothesize that the perception of sincerity and deceptive intent carry certain conceptual relation in their behavioral manifestations, and therefore resources in one domain can enhance performance in the other domain. We test multiple TL methods for enhancing sincerity and deception prediction on two datasets, with domain adaptation methods to mitigate the differences between the datasets. Our experiments target the overarching goal of enhanced human behavior modeling by sharing resources across various related problems and possibly learning a joint representation for various behavioral attributes.

Several researchers have investigated Transfer learning approaches [1], with one of the earliest explorations by Pratt [7]. More recent TL approaches include TL via dimensionality reduction [8], using deep learning representations [9, 10] and transfer learning in a transductive setting [11]. TL methods have been applied to sentiment classification [4], image classification [12] and other natural language understanding problems [13, 14, 15]. Recently, TL methods have been applied to modeling human behavior such as emotion recognition [15], facial expression recognition [16] as well as analysis [17].

Within the domain of sincerity and deception prediction, several modeling schemes were presented as part of the INTERSPEECH 2016 Computational Paralinguistics Challenge [18], including DNN rank learning [19], learning using phone recognition based features [20] and fusion of acoustic feature representations [21]. As a part of the challenge, Zhang et al. [22] proposed a transfer learning approach with multi-task learning perspective. Their proposal involves label generation for the task at hand from out-of-domain unlabeled data and retraining the model, a methodology also explored as a baseline in our experiments. Moreover, Zhang et al. [22] modify the continuous sincerity scores into binary labels for label compatibility. In our work, we retain the original sincerity and deception labels, thereby exploring methods that extend to labels with different annotation protocols. Additionally, we mitigate the differences between the dataset recording conditions using domain transfer techniques and incorporate them in our models.

Research in psychology has investigated the two concepts of sincerity and deception in various contexts such as integrity [23], leadership [24] and interpersonal perceptions [25]. Often these concepts are seen to oppose each other with low sincerity implying a deceptive intent. We test our models based on this hypothesis on a Deceptive Speech Database [18] and a Sincerity Speech Corpus [18], collected under different recording conditions and annotated by independent annotators. Initially, we train baseline models from in-domain resources without using any TL from the other dataset. This is followed by testing various settings of TL categorized as: (i) using datapoints from the other dataset and, (ii) appending sincerity/deception predictions from the other model as features. Finally, we also combine the two approaches and test using datapoints from the other dataset as well as sincerity/deception predictions from other model as features. The best models achieve an Unweighted Average Recall of 72.02% (baseline: 69.64%) on classifying deceptive vs non-deceptive intent and Spearman’s/Pearson’s correlation coefficients of 49.37%/48.52% (baseline: 46.51%/41.58%) between true and predicted sincerity scores. Finally, we analyze the results and present our conclusions. Next, we introduce the datasets followed by the methodology description in Section 3.

2. Datasets
We use two datasets for our experiments: (i) the Deceptive Speech Dataset and, (ii) the Sincerity Speech Corpus. Our models are evaluated using a cross-validation framework on the dataset portion with the availability of labels. We briefly describe the datasets below.

Deceptive Speech Dataset (DSD): The DSD dataset is available from the University of Arizona and we use a set of 1059 speech samples obtained from 49 participants. In each of
these samples, the participant either lies about his identity and behavior or tells the truth. Therefore, each of these samples is associated with a binary label of the speech being deceptive or not. 747 utterances (out of 1059) are marked as non-deceptive in the dataset. We refer the reader to [18] for further details.

**Sincerity Speech Corpus (SSC):** We use the SSC dataset provided by the Columbia University, consisting of 655 speech samples from 22 speakers. Unlike the labeling schemes for the DSD dataset, the samples in the SSC dataset are rated for perceived sincerity by a group of 13 annotators in a range of 0-4. The scores for each annotator were further normalized to zero mean and unit standard deviation. The final sincerity score for each sample is computed as the mean of thus normalized sincerity score from each annotator. Further details regarding the dataset and the annotation protocol are available in [18].

The goal of our experiments is to utilize the expected inherent relationship between deception and perceived sincerity labels. Therefore, in order to improve the performance for one dataset, we aim to utilize the knowledge from the other dataset by either directly using the data samples from that dataset or using the model learnt on the other dataset. However, ensuring the maximal transfer of knowledge from one dataset to the others is not trivial due to several dissimilarities between the datasets. These dissimilarities exist in the form of differences in dataset collection conditions and protocols, annotation procedures and the nature of labels. To begin with, we do not have any indication of (continuous) sincerity labels on DSD dataset and (binary) deception labels on the SSC dataset. Also, the speech samples in the DSD dataset are more similar to spontaneous speech, as the subjects were allowed to arrange speech and show emotions freely. In contrast, participants for the SSC dataset were asked to utter pre-specified sentences. We address these issues using a few label generation and data transformation techniques. The label generation and data transformation are performed using features extracted on each of the two datasets, discussed in the next section.

2.1. Features

We use a common set of 6373 acoustic features on the DSD and SSC datasets, termed as the ComParE Acoustic Feature Set [26]. These features are statistics computed on energy related descriptors (e.g. loudness, RMS energy, zero crossing rate), spectral descriptors (e.g. Mel Frequency Cepstral Coefficients, spectral energy) and voicing related descriptors (e.g. F0, voicing probability). Further details on these features can be obtained from [26].

3. Methodology

The goal of our experiments is to exploit the information from one dataset for improved target label prediction on the other dataset. For the purpose of this demonstration, we initially set a baseline on each of the two datasets. We then conduct two sets of TL experiments to utilize the information from a given dataset in improving performance for the other dataset. Below, we describe each of these modeling schemes in detail.

3.1. Baseline experiments

In the baseline systems, we only use the datapoints corresponding to each individual dataset, without any transfer of knowledge between the SSC and DSD datasets. The setup of the baseline experiments is similar to the one presented in the Interspeech 2016 challenge paper [18]. We describe the choice of baseline modeling schemes for the deception and sincerity prediction below.

**Deception prediction:** We train a Support Vector Machine (SVM) classifier on the ComParE Acoustic Feature Set [26] to predict the binary label of an utterance being deceptive or not. For evaluation purposes, we perform a 10 fold cross-validation, with each fold containing an independent set of speakers: 8 partitions are used as training set, 1 as development set and 1 as testing set. SVM parameters such as kernel and box-constraint are tuned on the development set. We use Unweighted Average Recall (UAR) as the evaluation metric on the DSD dataset, as was also the case during the Interspeech challenge 2016 [18]. UAR assigns equal weight to each class during recall computation, a case helpful when the class distribution is biased. The baseline results are listed in Table 1. Note that the results are slightly different than the one presented in the Interspeech 2016 challenge paper [18] as their evaluation defined a different partitioning scheme for the development and the testing set.

**Sincerity prediction:** Since the sincerity prediction involves continuous labels, we train a Support Vector Regressor (SVR) as the predictor. We use both Spearman’s and Pearson’s correlation evaluation metrics to obtain a sense of relative ordering as well as linear correspondence between predicted and true labels. In the baseline experiment, we perform a finer leave one speaker out cross-validation due to a small number of datasamples. Apart from the speaker in the testing set, other speakers are roughly equally divided between the training and the development set. The parameters for the SVR (kernel and box-constraint) are tuned on the development set. Table 1 presents the baseline results. The results are slightly different from the ones presented in the Interspeech 2016 challenge paper [18] due to differences in data partitioning and the fact that in [18], the box-constraint parameter was tuned globally for each fold. We, on the other hand, find the best parameters for each fold independently based on the development set.

3.2. Transfer Learning (TL)

Using TL, we aim to leverage one dataset for an improved performance on the other dataset. We conduct two TL experiments, namely: (i) adding datapoints from the other dataset during training and, (ii) appending sincerity/deception predictions from the other model as features. In the first approach, we expect that adding datapoints from the other dataset with synthetically created labels provides a lower generalization error [28].

In the second approach, we use sincerity predictions to improve classification of deception prediction (likewise using deception prediction for sincerity scoring). We describe the experiments in detail below.

3.2.1. TL: adding datapoints from the other dataset during training

In this section, we append the datapoints from the other dataset in order to improve the performance for the task of interest. First, we generate synthetic labels for the task at hand. We then conduct two TL experiments to utilize the information from a given dataset in improving performance for the other dataset. Below, we describe each of these modeling schemes in detail.

(a) Random Sample Consensus (RANSAC): During the RANSAC [29] implementation, an initial model is trained on the available dataset with labels. Then, the algorithm performs prediction on the datapoints from the other dataset without the labels of interest. A fraction of these datapoints is then added to the existing labeled dataset and an updated model is trained with the additional labeled data. Researchers have proposed several criteria for selecting the fraction of data (e.g. adding the datapoints farthest away from class boundaries). This procedure is repeated iteratively till a certain criterion is met (in our experiments, we perform the iterations till the performance
on the held out development set is maximized). Studies have investigated the use of RANSAC in several applications including image analysis [29] and computer graphics [30]. Since the task objective is different for the DSD and the SSC corpus, we describe them separately below.

Deception prediction: For the purpose of deception prediction, synthetic labels are obtained on the SSC dataset. We use the same train, test and development set split as in the baseline experiments. Initially, an SVM model is obtained using the train partition of the DSD dataset and datapoints are added from the SSC dataset till the performance on the development partition of the DSD dataset is maximized. In each iteration, datapoints farthest away from the class boundary are added [31]. We tune the proportion of SSC data added at each iteration from farthest 5% to all of the data.

Sincerity prediction: The details of sincerity prediction are similar to deception prediction, except for the fact that the model we train is an SVR. We implement the regression version of RANSAC algorithm as proposed in [32]. In this algorithm, we iteratively add a cluster of points with least prediction error (as per the existing regression model) to the training data. The train, test and development set partitions are kept the same as the baseline experiments. Table 1 shows the results for the RANSAC algorithm implementation.

(b) Label transformation: The RANSAC algorithm does not make use of sincerity labels during training deception models (and vice versa). Rather, new labels are generated using machine learning models. Alternatively in the label transformation method, we transform the continuous sincerity labels into binary deception labels while training deception models (and vice versa while training sincerity models). We describe the label transformation for deception and sincerity prediction below.

Deception prediction: In this experiment, we append a part of the sincerity dataset to the DSD dataset during model training. The portion of the sincerity dataset with perceived sincerity below (above) a certain threshold are marked as deceptive (non-deceptive). These thresholds are tuned as a pair for maximal performance on the development partition of the DSD dataset.

Sincerity prediction: Obtaining sincerity labels on the deception dataset involves transforming binary labels into a continuous sincerity scale. For our experiments, we approximate all deceptive utterances to carry a fixed sincerity rating. Similarly, all non-deceptive utterances are labeled with a different constant sincerity rating. These ratings for the deceptive and non-deceptive utterances are again tuned as a pair for maximal performance on the development partition of the SSC dataset.

The results for the label transformation are listed in Table 1.

3.2.2. TL: using sincerity/deception predictions as features

We next propose using sincerity prediction as features during deception prediction and vice versa. In these experiments, we examine the hypothesis that obtaining a sincerity score would be useful for deception prediction and vice versa, unlike obtaining same label on additional data as explored in previous section 3.2.1. In the experiment, we predict the sincerity scores (/deception prediction) on the DSD (SSC) dataset to aid deception (/sincerity) prediction. These sincerity (/deception) scores are obtained from a model trained on an adapted version of the SSC (DSD) dataset. As there exists a mismatch between the SSC and DSD datasets, we perform a domain adaptation to reduce the differences in the distribution of SSC and DSD datasets. We provide the experimental details for the experiments below.

Deception prediction: We use the train, test and development set partitions for the DSD dataset as described before. In order to obtain the sincerity scores on the DSD dataset, we train a model on an adapted version of the SSC dataset. Through adaptation, we aim to mitigate the differences between the two datasets in terms of the data recording conditions. We use the Geodesic Kernel Flow (GFK) adaptation [33] to this effect. The GFK transformation treats the DSD dataset as the source dataset and adapt it to have a similar distribution like the DSD dataset in a shared subspace (which is different from the original feature subspace). We then train an SVR model, SVR S-GFK on the SSC dataset and make predictions on the DSD dataset, in the subspace learnt by GFK transformation. For final deception prediction, the sincerity scores obtained from SVR S-GFK are used in conjunction with the original acoustic features (in Section 2.1) in a stacked generalized framework [34]. Figure 1(A) depicts a summary of this stacked generalization framework. In the stacked generalized framework an SVM model, SVM D1 is first trained using the original acoustic features from the DSD training set (this model is same as the baseline deception prediction model). Then, applying SVM D1, we obtain the distances from SVM decision hyperplanes as soft estimate of the deception scores. We then train a final SVM model, SVM D2 that combines these soft deception scores from SVM D1 and the sincerity scores from SVR S-GFK. SVM D2 model is trained on the combined training and development partitions of DSD dataset to prevent overfitting to the training set.

Sincerity prediction: Akin to the deception prediction experiments, we use the predefined train, test and development set partitions for the SSC dataset. In order to obtain deception predictions on the SSC dataset, we train an SVM model, SVM S-GFK on an adapted version of the DSD dataset. The adaptation is again performed using the GFK transformation, with DSD dataset modified as source dataset to resemble SSC dataset as the target dataset. Instead of using the binary deception labels from SVM D-GFK on SSC dataset, we again use distance of datapoints from SVM decision hyperplanes as a soft deception score estimate. We then train a first layer SVM model (SVR S1), trained on the SSC training partition, predicts the sincerity scores based on the original features (this model is same as the baseline sincerity prediction model). A second layer SVM model (SVR S2) then combines the predictions from SVR S1 with the deception scores from SVM D-GFK to provide the final sincerity scores. SVM S2 is trained on the concatenated train and development sets to prevent overfitting to the training set. Figure 1(B) summarizes the stacked generalization model for sincerity prediction. In the next section, we combined the adding datapoints and appending sincerity/deception outcomes as features.

3.2.3. TL: adding datapoints + using sincerity/deception predictions as features

As the next step, we combine the previous two TL approaches of adding datapoints and sincerity/deception predictions as features. We describe this approach below.

Deception prediction: This experiment is similar to the deception prediction in the last section involving the incorporation of sincerity/deception scores as features during prediction. The only difference is in the first stage of stacked generalization where we also append the SCC dataset to the DSD training partition to learn SVM D1. SVM D1 could be trained using either RANSAC or label transformation. The second stage SVM, SVM D2 is then trained on the DSD training and development partitions using the hyperplane distances predicted from SVM D1 and the sincerity scores predicted by SVR S-GFK. The final predictions on the DSD testing set are made by SVM D2.

Sincerity prediction: We modify the first stage of stacked generalization as presented for sincerity prediction in Section 3.2.2 by adding the DSD dataset to the SSC training partition in learning SVR S1. SVM S1 is trained either using RANSAC
or label transformation. In the second stage of stacked generalization, SVR$_{S2}$ is then trained on combined training and development sets with outcomes from SVR$_{S1}$ and SVM$_D$ as features. We again use distances from SVM$_D$ as features as soft estimates of deception labels. SVR$_{S2}$ provides the final decisions on the SSC testing set.

In the next section, we present our results for the baseline and various TL models and present discussion.

## 4. Results and Discussion

Table 1 shows the results for the various deception and sincerity prediction experiments. For the last TL experiment involving both adding of datapoints and using sincerity/deception predictions as features, we perform an additional experiment where the method for training SVM$_{D2}$ and SVR$_{S1}$ (RANSAC/Label transformation/ not adding other dataset) for each cross-validation fold is chosen based on their performances on the development set. We perform this experiment to get a fair assessment of the TL on the test set, instead of advocating for a TL approach solely based on the results observed on the test set (a scenario not possible in the real world).

**Discussion:** From the results in Table 1, we observe that the additional information provided in the other datasets has minimal positive impact, with only RANSAC showing improvements in case of deception prediction. This may stem from the fact that the labels are synthetically generated, as well as that the distribution of the data samples in DSD and SSC datasets are not similar due to difference in recording conditions and data collection protocols. Therefore, adding the other dataset with synthetic labels only serves as a source of noise. On the other hand, adding sincerity/deception scores after domain adaptation improves over the baseline. The UAR improvement is significant in case of deception prediction using McNemar test (p<5%) modified for UAR, accounting for data imbalance [35]. The improvement in Spearman’s correlation for SSC dataset is not significant using Fisher r-to-z transformation, although the Pearson’s correlation does show a marginally significant improvement (Fisher r-to-z transformation, p<10%). Finally, adding datapoints from other dataset along with using sincerity/deception scores as features again gives mixed results with improvements only for label transformation. We are encouraged to see that the final assessment on the test set by tuning the method shows improvement for both the tasks over the baseline. Although the performance for deception detection is worse off than only using model prediction as features, tuning provides the best overall performance for the sincerity task. Additionally, we obtained the sincerity prediction on an additional test set provided the Interspeech challenge 2016 on the Sincerity corpus, consisting 256 of utterances [18]. Our best system on the development data obtains a ρ$_S$ of 61.8% on this test set, outperforming the baseline value (ρ$_S$ = 60.2%) presented in [18]. In particular, we observe that using sincerity/deception scores as features during prediction delivers promising results.

On computing the absolute Spearman’s correlation coefficient values between the predictions from SVM$_D$ and the true deception labels, we obtain a value of 19.1%. Similarly the absolute correlation coefficient between the deception scores obtained from SVM$_D$ and the true sincerity scores is 20.2%. Although these correlation coefficients are modest, they show meaningful correlation trend with deceptive labels having a low sincerity score and vice versa (in both the transformation experiments). Adding datapoints from the other dataset does not always improve the performance, therefore indicating that the synthetic label generation needs improvements. One future work in this regard is addressing the difference between the datasets while generating synthetic labels.

| Table 1: Results for predicting binary deception labels and sincerity scores on the DSD and SSC datasets, respectively. We report Unweighted Average Recall (UAR) for DSD dataset and Spearman’s/Pearson’s correlation coefficients (ρ/ρ$_p$) for the SSC dataset. Best performances are marked in bold. |
|-----------------|-----------------|-----------------|
|                 | Deception       | Sincerity       |
|                 | (UAR in %)      | (ρ, ρ$_p$ in %) |
| Baseline        | 69.64           | 46.51, 41.58    |
| TL: adding datapoints from other dataset |
| RANSAC          | 70.26           | 48.45, 33.66    |
| Label Transformation | 67.17           | 43.44, 29.36    |
| TL: Sinc./Dep. model prediction as features | 72.02 | 48.53, 47.84 |
| + RANSAC        |                | 48.50, 48.04    |
| + Label transformation | 69.90           | 49.37, 48.18    |
| + Tuned         | 70.59           | 49.37, 48.52    |

## 5. Conclusion

Transfer Learning has shown promise in several domains including human behavior modeling, where it can be expensive and time prohibitive to obtain large amounts of appropriately labeled data. In this work, we apply TL methods to the problems of sincerity and deception prediction using behavioral data from two distinct datasets; the DSD and SSC datasets. We test two methods: (i) adding datapoints from the other dataset and, (ii) appending system predictions from the other tasks as features. We also test a combination of the two methods and present our analysis. We obtain superior performances to models trained only on in-domain data, indicating the promise of TL methods in such problems.

In the future, we aim to explore other methodologies within the domain of TL to the problem of sincerity and deception prediction (e.g. other domain adaptation methods [36]). One could also test the applied methods to other human behavior modeling domains such as engagement [37] and interestingness [38]. Finally, we are also working on formulating a theoretical understanding of the transfer of knowledge between correlated concepts with differences in data recording conditions.

---

1. Sincerity scores as features during prediction delivers promising results. Where the method for training SVMD$_1$ and SVRS$_1$ (RANSAC/Label transformation/ not adding other dataset) for each cross-validation fold is chosen based on their performances on the development set.

2. We also obtained the sincerity prediction on an additional test set provided the Interspeech challenge 2016 on the Sincerity corpus, consisting 256 of utterances [18]. Our best system on the development data obtains a ρ$_S$ of 61.8% on this test set, outperforming the baseline value (ρ$_S$ = 60.2%) presented in [18]. In particular, we observe that using sincerity/deception scores as features during prediction delivers promising results.

3. On computing the absolute Spearman’s correlation coefficient values between the predictions from SVM$_D$ and the true deception labels, we obtain a value of 19.1%. Similarly the absolute correlation coefficient between the deception scores obtained from SVM$_D$ and the true sincerity scores is 20.2%. Although these correlation coefficients are modest, they show meaningful correlation trend with deceptive labels having a low sincerity score and vice versa (in both the transformation experiments). Adding datapoints from the other dataset does not always improve the performance, therefore indicating that the synthetic label generation needs improvements. One future work in this regard is addressing the difference between the datasets while generating synthetic labels.
6. References


