Combining vs. Transferring Knowledge: Investigating Strategies for Improving Demographic Inference in Low Resource Settings

Yaguang Liu
Georgetown University
Washington, USA
yl947@georgetown.edu

Lisa Singh
Georgetown University
Washington, USA
lisa.singh@georgetown.edu

ABSTRACT

For some learning tasks, generating a large labeled data set is impractical. Demographic inference using social media data is one such task. While different strategies have been proposed to mitigate this challenge, including transfer learning, data augmentation, and data combination, they have not been explored for the task of user level demographic inference using social media data. This paper explores two of these strategies: data combination and transfer learning. First, we combine labeled training data from multiple data sets of similar size to understand when the combination is valuable and when it is not. Using data set distance, we quantify the relationship between our data sets to help explain the performance of the combination strategy. Then, we consider supervised transfer learning, where we pretrain a model on a larger labeled data set, fine-tune the model on smaller data sets, and incorporate regularization as part of the transfer learning process. We empirically show the strengths and limitations of the proposed techniques on multiple Twitter data sets.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

demographic inference, transfer learning, data set distance, CLIP

ACM Reference Format:
Yaguang Liu and Lisa Singh. 2023. Combining vs. Transferring Knowledge: Investigating Strategies for Improving Demographic Inference in Low Resource Settings. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570462

1 INTRODUCTION

It is well known that many deep learning methods need large amounts of labeled data to perform well on different text classification tasks [28]. However, there are many tasks for which the amount of available training data is limited because of the high cost of labeling or the impracticality of finding a large number of examples. Demographic inference on social media is one such task.

There are many reasons the task of demographic inference is important. First, when determining the fairness of an algorithm, we need demographic characteristics of users, even for data sets where they may only be available for a subset of the users. Also, users of social media platforms, such as Twitter, share their activities, interests, and beliefs with others on the platform. These pieces of information are potentially important insights for social science and policy researchers studying the evolving public opinion on issues of the day, particularly as survey response rates continue to decline. In order for researchers to effectively use public social media platforms to understand public opinion, they need basic demographic information to identify subgroups within their samples, and in some cases, to ensure a representative sample of individuals [4]. However, these demographic attributes are not always explicitly shared by users on social media through their profiles or metadata, thereby requiring machine learning to infer these characteristics. Demographic inference is a user-level task where multiple posts (tweets) are used to infer the demographic of interest.

For most user demographics, manual labeling is required to determine the demographic, leading to very small training data sets, ranging from 100s to 1000s of users [5, 23]. Therefore, in this paper, we explore different strategies for demographic inference on Twitter in a constrained environment, i.e., when the amount of labeled data is limited. Despite the recent attention to the demographic inference task [5, 14, 16, 24], work remains to address the low resource constraint [17]. This paper focuses on inferring two well studied demographic attributes on Twitter: gender and age.

Different strategies have been proposed to tackle the low resource constraint for other inference tasks, including transfer learning, data augmentation, and data combination [3, 18, 32]. We investigate two of these approaches. First, we consider the scenario when training data sets are of a similar size. In this case, we explore data combination – a strategy that merges labeled data sets of similar size prior to training. We refer to this strategy as demographic inference data combination (DIDC), and show that it is reasonable when data set distributions are similar to each other. To evaluate such similarity between data sets, we propose using a variant of the optimal transport computation [2]. Next, we consider the scenario when the labeled training data sets are of different sizes. In this case, we implement demographic inference transfer learning (DITL) by training on a large labeled data set, fine-tuning the model on the smaller-scale data sets, and incorporating regularization as part of the transfer learning process. Our experiments demonstrate that DITL can generally improve the performance, regardless of the level of similarity between the data sets.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570462

Finally, most research [13, 14] on age and gender inference focused on text analysis uses Bidirectional Encoder Representations from Transformers (BERT) [7] as the auxiliary embedding space to improve predictive accuracy. In this work, we also explore adopting an embedding space that combines both visual and text knowledge. We use Contrastive Language-Image Pre-Training (CLIP) [21], to learn visual concepts from images and language concepts from text. By considering both BERT and CLIP, we can also answer the question - does an embedding space built using just text vs. one built using both text and images matter for demographic inference?

In summary, **this paper makes the following contributions**. (1) We propose DIDC, a method that combines small demographic inference data sets of similar magnitude, and analyze the combination using a variant of the optimal transport computation. (2) We propose DITL to help when data sets vary in size and overfitting becomes a larger issue. (3) We explore different auxiliary embedding spaces when generating our model. (4) We empirically evaluate DITL and demonstrate its effectiveness across different data sets. (5) We make our labeled data set and code publicly available.[1]

2 RELATED LITERATURE

This section begins by discussing the demographic inference literature. Then, we review the previous literature about classifying small datasets. Finally, we present research that considers different ways to compute data set distribution distance.

Demographic Inference Most early research on demographic inference used classic algorithms such as logistic regression (LR), support vector machines (SVM), and random forest (RF), usually with bag of words as features [5, 13, 14, 20, 24, 30]. Some papers investigated using stylistic features such as punctuation [14]. In recent years, research has shifted to using deep learning models. Miura and colleagues propose a gated recurrent unit (GRU) model [6] for location inference that combines tweet text, biography and network data using an attention mechanism [16]. Liu and colleagues [14] develop a fine-tuned BERT model, pretrained using a Siamese network for gender and age prediction. A BERT emoji model that takes advantage of a hierarchical architecture is proposed by Liu and Singh [13]. Their work uses a GRU with an attention layer to separately train the emoji component (using word embeddings and a Convolutional Neural Network - CNN) and the text component (using BERT). However, all these neural models require larger amounts of labeled data (10,000s to 100,000s). Klein and colleagues [10] propose ReportAGE that extracts the age of Twitter users from tweets using regular expressions. But their method reflects a selection bias toward younger users who are possibly more willing to reveal their age information. Their coverage (around 54%) also suggests that this approach is useful for opportunistic sampling, but less reliable for broader population sampling.

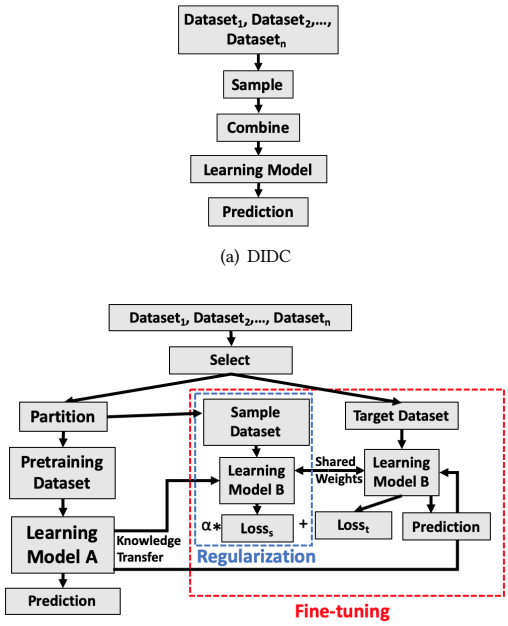
Some approaches for demographic inference consider using profile images in the training data. Vijayaraghavan and colleagues [30] use the inception architecture [29] to extract features from images. Wang and colleagues [31] propose a multi-modal model using profile image, username and biography. They use DenseNet to map images into an image embedding space. Different from previous work, we investigate combining multiple smaller training data sets or using transfer learning to improve the quality of the model. Similar to a few of the earlier works, we focus on using only post content for this inference task because collecting biographies, user networks and/or images through the Twitter API can be costly. Finally, this work is the first to compare the use of a text and an image/text auxiliary embedding spaces within these models. To the best of our knowledge, in general, there is not a paper that focuses on demographic inference within a constrained environment.

Algorithms for Small Data Models developed on small data sets suffer from overfitting and therefore, do not generalize well. Different methods have been proposed to address this issue. A simple technique is to add a regularization term on the norm of the weights for models such as logistic regression [19]. Dropout, which works by probabilistically removing neurons from designated layers during training, has also been shown to work well to prevent overfitting [27]. Some research introduces data augmentation to boost the performance of small data sets. For example, Wei and Zou [32] propose Easy Data Augmentation (EDA) and show improvement by augmenting data with techniques such as synonym replacement. However, the main limitation of data augmentation arises from the data bias, i.e., the distribution of the augmented data could be very different from the original data [33].

In recent years, transfer learning has become an important approach for improving deep learning on small data sets. These approaches often employ supervised pretraining to transfer knowledge between related domains. For example, Mou et al. [18] show that by training on a large data set using a recurrent neural network (RNN) model, and then fine-tuning the model on a smaller data set, the performance on a binary sentiment task improved by approximately 6%. Semwal and colleagues [25] propose using CNN and show that transfer learning from a supervised model to another data set can be helpful for many text classification tasks (not demographic inference). These previous works focus on post-level classification. Shang et al. propose a transfer learning model for demographic inference, but their approach is different from traditional transfer learning via user modeling since their task has no available demographic information in the target domain. In their paper, they introduce a model based on matrix factorization and the only feature used is users' ratings for a movie or book [26]. To the best of our knowledge, this work is the first to study supervised transfer learning for demographic inference on social media.

Data Set Distribution Distance Researchers have introduced numerous algorithms to compute data set distribution distance [1, 2, 34]. For example, Achille et al. [1] use the Fisher information metric to construct vector representations of data sets which they then use to define the data set similarity. Another approach is optimal transport. Its goal is to look for a transport map that can be used to transform one probability density function into another. Yurochkin and colleagues use optimal transport distance for document similarity, defining an inner-level distance between topics and an outer-level distance between documents [34]. Alvarez-Melis and Fusi [2] extend the work and propose Optimal Transport Dataset Distance (OTDD). They compute data set distance using optimal transport, where a data set contains many samples, and they consider data set labels and show their relevance within the data set similarity computation. They also show that data set distance is

[1] https://github.com/GU-DataLab/Demographic-Inference



(a) DIDC

(b) DITL

Figure 1: Overview of the proposed approaches

highly correlated with the transferability between data sets. Because our prediction task is a user level task, we will extend this previous literature by computing user level data set distance, not just post level.

3 METHODS

This section describes our proposed approaches for dealing with the resource constraint. Figure 1 shows the high-level processes. The process of DIDC (Section 3.1) is illustrated in Figure 1(a) and Figure 1(b) shows the procedure for DITL (Section 3.2).

3.1 DIDC

3.1.1 Data Construction. It is not unusual for social science researchers to have access to a few small labeled training sets that are independently collected and labeled. However, because of the training data size, the predictive power is low for neural models. Therefore, we explore combining training data sets of similar size. In cases where the sizes are not similar, we combine samples of similar size to avoid biasing the combined training data. Formally, suppose we have access to n data sets, ranging in size $|d_1|, |d_2|, ..., |d_n|$. Let's also suppose that we have labeled training data for each of these n data sets. We take a sample $s_1, s_2, ..., s_n$ from the n training sets, respectively, where $|s_1| \approx |s_2| \approx ... \approx |s_n|$ and construct a new labeled training set (T) that contains these samples – $T = \cup_{i=1}^{n} s_i$. Although the merged training set T may still be small, it is approximately n times as big as each individual sample training set, when s_i is of similar size to d_i.

3.1.2 Data Set Distribution Similarity. To better understand when and how multiple data sets should be combined, researchers have proposed using data set similarity [35]. Here we propose using user-level OTDD to compute the data set distribution similarity and then quantify the performance of the combination.

OTDD Using optimal transport to compare two probability distributions requires defining a distance between points sampled from those distributions. For OTDD, when comparing two data sets, each point is a pair consisting of the feature representation and the label. The feature representation can be generated by mapping a post to an embedding space using a pretrained model, such as BERT. Labels are represented as conditional probabilities: $P_y = P(X|Y = y)$, where X is the feature set and y is the label of a data point. Thus, the distance of two users (data points) is as follows:

$$d(z_1, z_2) = (d(x_1, x_2)^2 + W(P_{y1}, P_{y2})^2)$$

where z represents a user, whose feature is expressed as a single post, W is the Wasserstein distance and P is the label representation. Once we have the distance between users, we can use it to determine the distance between distributions over feature-label pairs (training set), which is the OTDD:

$$OTDD(D_1, D_2) = min_{\pi \in \prod(P_1, P_2)} \int_{z \times z} d(z_1, z_2) d\pi(z_1, z_2)$$

where D represents a training set, π is a joint distribution (formally, a coupling) with marginals P_1 and P_2. Although OTDD has shown strong performance in measuring the distance between post-level data sets, our task contains user level training data. Therefore, we introduce the user-level OTDD to compute the data set distribution distance for demographic inference using multiple posts. Specifically, we represent a user by the mean value of all of her/his embeddings, and then use the same process as post-level OTDD to get the data set distribution distance. As our empirical evaluation will show, though simple, this metric is very effective in measuring the distance between user-level data sets.

3.2 DITL

We focus on the simplest approach for transfer learning in this paper (see Figure 1(b)) and leave techniques like layer freezing for future work. We train the model and fine-tune it directly for a new data set. We then apply regularization to further avoid overfitting.

As pointed out by Zhou et al. [36], one weakness of only performing the source task and then the target task directly is that the text encoder learned in the original task may be overridden after being fine-tuned in the target task. If the target training set is too small, the fine-tuned encoder has a high risk of overfitting the target data. This is the main motivation for introducing an additional component to regularize the model. Specifically, as shown in Figure 1(b), we first select the largest data set represented as L, and then take out a sample data set M that has a similar size as the target data set T, i.e., $|T| \approx |M|$ (This is always doable as $|L| > |T|$). Next, we pretrain our model using the remaining data P (pretraining data), where $|P| + |M| = |L|$. We then fine-tune the pretrained model using the target data set T and the sample data set M simultaneously. This additional component, using M as input, serves as a regularizer. The basic idea is as follows: by making the model train on the similar type of data as in pretraining (as they are from the same data set), there will be less influence of the smaller data set on the final model. Finally, we minimize the sum of two losses from the regularizer

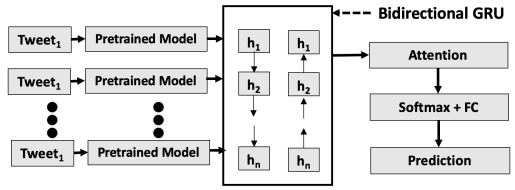


Figure 2: Overview of the proposed learning model

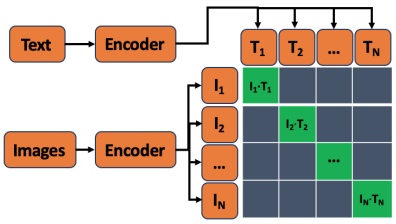


Figure 3: Contrastive Pretraining

and the target task: $loss = -(\alpha Y_s \cdot log \boldsymbol{p}_s + Y_t \cdot log \boldsymbol{p}_t)$ where Y_s and Y_t are one-hot encodings of the source and target data set labels, $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$ are the prediction values, and α is a weight to control the importance of regularization.

3.3 Learning Model

The learning model we use is similar to prior work (see Figure 2). Specifically, tweet text is first input into the pretrained model (BERT or CLIP) to compute embeddings. Then the vectors are input into an RNN with attention. We now present the details of our model.

BERT BERT is a deep neural model that processes text bidirectionally (from left to right and right to left). BERT uses an encoder stack transformer architecture. This architecture is an encoder-decoder network using self-attention. Using a large amount of text data, it is pre-trained on two tasks: Masked Language Modeling and Next Sentence Prediction.

CLIP CLIP is a deep neural model that is constructed using both text and image data. As shown in Figure 3, the model tries to learn the relationship between an entire sentence and the image it describes with the goal of maximizing the similarity of the diagonal (the green area - $(I_1T_1, I_2T_2, ..., I_NT_N)$) and minimizing the remaining area. Prior research has shown that CLIP is capable of predicting the most relevant text snippet given an image, and conducting image classification in computer vision with zero-shot capabilities [21].

Learning Model Structure For the text representation, we use the pretrained model directly to generate the embeddings, having tweet text as input. We use a GRU structure to encode all the tweet text for each user. Next, we adopt an attention mechanism so that the model is able to selectively focus on valuable parts of the input text for our task and learn the association between them. Finally, we send this output into a Softmax layer for the final prediction.

4 EMPIRICAL EVALUATION

In this section, we begin by introducing the data sets we use (Section 4.1) and the experimental setup (Section 4.2). We then present an empirical evaluation of the effectiveness of our two approaches (DIDC and DITL) in a low resource setting (Sections 4.3 and 4.4). We also explore different pretrained models for demographic inference, and provide some intuition on the impact of different embedding spaces for modeling Twitter language (Section 4.5). Finally, we compare user-level and post-level inference (Section 4.6).

Demographics		Category	Count			
			Wiki	IMDB	Survey	Merged Data
Gender	-	Female	3335	1454	289	720
	-	Male	7891	1911	383	788
Age	Bin2	<45	7538	1898	324	787
		>=45	3731	1467	348	721
	Bin3	<35	5206	807	178	465
		35-54	3907	2013	296	592
		>=55	2156	545	198	451
	Bin4	<30	4038	388	-	-
		30-40	4254	962	-	-
		40-50	2340	1096	-	-
		>=50	3683	919	-	-

Table 1: Ground truth data distribution

4.1 Data Sets

Wiki We use the Wikidata benchmark data set from Liu et al. [14]. It contains a mapping between user demographics (gender and age) and Twitter handles. We use the Twitter API to retrieve users' most recent posts. The average number of tweets per user is 160.

IMDB Beginning with a public data set containing the demographic information of actors and actresses in IMDB, we use different celebrity lists to identify the Twitter handles of different celebrities. Following that, we collect the tweets of each celebrity using the Twitter API.[2] The average number of tweets per user is 187. There are no overlapping handles between the Wiki data and the IMDB data.

Survey Data Our research team conducted a nationally representative survey related to Covid-19. Those respondents who used Twitter were also asked if they would consent to allow our research team to download their tweets for both computer science and social science research. This data set contains those who consented.[3] The average number of tweets per user is 149.

Merged The merged data set is a combination of data from different sources. We use the entire Survey data set, and randomly sample similar numbers of users from the Wiki and the IMDB data sets.

For all the data sets, we follow the same pre-processing procedure: (1) remove users that have less than 20 English tweets, (2) lowercase all the words, and remove stopwords, handles and mentions for the classic machine learning models.

Table 1 shows the number of users in each data set for gender and age category. Because of training data limitations, we consider a binary version for gender. For age, we consider a binary task with 2 age bins and a multi-class version with 3 and 4 bins for the Wiki and the IMDB data set, and 3-bins only for other smaller data sets. Forty-five is viewed as a new era of adulthood according to the

[2]The data can be found at https://github.com/GU-DataLab/Demographic-Inference/tree/main/dataset.

[3]This research was approved by Georgetown University's Institutional Review Board (number: STUDY00002133).

Levinson adult development model [12]. Thus, we choose 45 as the 2-bin dividing line. The 3-bin and 4-bin boundaries were identified by social science researchers on our team.

Baseline Models: The classic models we compare to are LR proposed by Nguyen et al. [20], SVM by Chen et al. [5], and RF [30]. The neural models we compare to are Vanilla BERT [14], Siamese BERT [14], and BERT emoji [13].[4]

4.2 Experimental Settings

We use two NVIDIA Tesla P4 GPUs, each having 16 GBs of memory. We adopt the CLIP base model. Its text encoder is a Transformer with a modified architecture [22], and the image encoder is the Vision Transformer (ViT-B/32) [8].[5] We use the Adam update rule [9] to optimize our model. Weight, bias, and context vector are randomly initialized for the attention layers and then normalized with a mean value of 0 and a standard deviation of 0.05. They are jointly learned during training. Gradients are clipped between -1 and 1. Batch size is set to 32. The initial value of α is set to 1 and the learning rate is set to 0.0001. The maximum number of tweets per user is 200 and those having fewer are padded. We randomly sample from the group using the Python library *imblearn* [11] in order to create more balanced data sets. We run each experiment ten times. Each time we divide the data into training, validation and test set with a random seed. We report the average F1 scores, as well as 0.95 confidence interval. For DITL, we first use 32 samples per class from the target set for fine-tuning and then all the target set data.[6]

4.3 Results Using DIDC

Our results for the proposed model and DIDC are presented in Table 2. Values shown in dark red are the highest scores. Values shown in red are the highest scores that are statically significant using a p-value of 0.05.

Wiki Analysis: Beginning with gender, the best classic model has an F1 score of 0.839 and the best neural model achieves an F1 score of 0.882. Our model using CLIP performs better than the state-of-the-art classic model and neural model by 5.9% and 1.6%, respectively. For age classification with 2-bin, we observe that the best classic model is RF, and it has an F1 score of 0.816. The best neural model achieves an F1 score of 0.824, which is only a marginal difference from the classic model. The proposed model using CLIP is again the best one and performs 3.3% better than the best previous model. The proposed model using BERT is generally not as strong.

For age with multiple-bins, RF has the highest results among classic models, but it is approximately 4% lower and 5% lower than the best neural model for age with 3-bin and 4-bin, respectively. For 3-bin, our model using CLIP achieves an F1 score of 0.72, which is 3.4% higher than the state-of-the-art model. For 4-bin, the proposed model using CLIP is 5.9% higher than the best neural model.

IMDB Analysis: For IMDB, the classic models perform similarly. The neural models are generally not as good as the classic models for both gender and age. The one exception is the proposed model using CLIP. We see that for gender, it has an F1 score that is 3.6% higher than the best state-of-the-art model. For binary age, this model achieves the highest F1 score, which is 3.2% higher than the best previous model. CLIP for age with 3 bins is 3% higher than the state-of-the-art and 5.8% higher for 4-bin. We again see that the proposed BERT model does not perform as well as the CLIP model.

Survey Analysis: Similar to the Wiki and IMDB data, our model using CLIP generally performs better than the state-of-the-art. However, there are a few things to note. First, we can see that for gender, the proposed model using CLIP has only a marginal difference when compared to the best previous model. For 2-bin age, it is also only 1.2% higher. We hypothesize that this is caused by the small size of the Survey data set and likely overfitting that is occurring, highlighting the challenges associated with small data sets.

Merged Data Analysis: Similar to our previous findings, the proposed model using CLIP performs better than the state of the art for both age and gender, ranging from 2.5% to 7.5%. The results across all these data sets suggest that our proposed model performs better than the state-of-the-art and that using text-image embeddings (CLIP) for pre-training is more robust for noisy tweets than using text embeddings (BERT) alone.

However, our main interest is in understanding whether a combined training data set leads to stronger results than using a smaller training data set on its own. Table 4 shows this comparison for our model using CLIP across the data sets.[7] It shows the F1 scores when the three sampled training sets are trained independently (TI) and when they are merged (TM). We see that the difference in F1 score using the Wiki data is marginal for gender and 3-bin age. For 2-bin age, the result is approximately 4% lower. The performance of both IMDB and Survey data sets improves significantly across almost all demographics. Although intuitively Wiki and IMDB seem more similar since both of them include famous individuals, the performance suggests that the two training sets do not supplement each other well.

To better understand this finding, in Table 3, we present the distribution distance between different data sets using user-level OTDD. We can see that the distance between the IMDB and Survey data sets is smaller than IMDB and Wiki or Survey and Wiki. This may also explain why the performance of TM for IMDB and Survey improves (compared to TI), while it decreases for the Wiki data set.

To further show the correctness of user-level OTDD, we adopt a similar approach as Alvarez and Fusi [2] to demonstrate the relationship between distance and learning performance. We simulate an adaptation setting. For every pair of labeled sets,[8] we first train the model using the entirety of the source domain data, after which we fine-tune and evaluate it on the target domain. For example, to test the transferability for gender inference from IMDB to Wiki, we will train the model using all the data from IMDB and then fine-tune it with limited samples from the Wiki data set.[9] Figure 4 shows that the data set distribution distance computed using user-level OTDD is highly correlated with transferability. Specifically, as the

[4]We explored using a transformer architecture, and also tried RoBERTa, but the difference is marginal.

[5]OpenAI has not released the data set yet.

[6]We have tried both 32 and 64 samples and while 64 gives higher F1 scores, we present the results using 32 samples since our environment is resource constrained.

[7]Due to space limitation, we focus on analyzing the distances based on CLIP as it performs better, but note that this can also be applied to BERT.

[8]We ensure that labeled sets are of a similar size by sampling Wiki and IMDB.

[9]We have done a sensitivity analysis by using both 32 and 64 users for fine-tuning and found that the curve is similar for both cases. We choose 64 for demonstration as it has a more balanced positive and negative increase in F1 score.

Model	Wiki				IMDB				Survey			Merged Data		
	Gender	Bin2	Bin3	Bin4	Gender	Bin2	Bin3	Bin4	Gender	Bin2	Bin3	Gender	Bin2	Bin3
Unigram-RF	0.812	0.816	0.649	0.518	0.835	0.717	0.583	0.445	0.68	0.663	0.509	0.751	0.741	0.559
Nguyen et al.	0.839	0.782	0.63	0.513	0.821	0.701	0.556	0.451	0.698	0.672	0.518	0.738	0.716	0.512
Chen et al.	0.816	0.797	0.635	0.509	0.831	0.722	0.587	0.454	0.656	0.663	0.504	0.738	0.74	0.543
Vanilla BERT	0.867	0.785	0.611	0.436	0.812	0.689	0.576	0.435	0.706	0.691	0.517	0.775	0.722	0.503
Siamese BERT	0.87	0.783	0.614	0.455	0.827	0.665	0.519	0.43	0.704	0.629	0.495	0.753	0.699	0.516
Liu et al.	0.882	0.824	0.686	0.569	0.828	0.697	0.561	0.452	0.679	0.721	0.555	0.775	0.757	0.558
Proposed BERT	0.868	0.817	0.679	0.55	0.813	0.688	0.558	0.456	0.678	0.708	0.563	0.762	0.745	0.572
Proposed CLIP	**0.898**	**0.857**	**0.72**	**0.628**	**0.871**	**0.754**	**0.617**	**0.512**	**0.712**	**0.739**	**0.629**	**0.811**	**0.782**	**0.634**

Table 2: F1 score for age and gender on different data sets. Bin# refers to the number of bins for age.

data sets	Gender	Bin2	Bin3
IMDB ⇔ Survey	5.84	5.86	5.9
WIKI ⇔ Survey	8.67	8.74	8.78
WIKI ⇔ IMDB	8.39	8.46	8.5

Table 3: Distance between two data sets

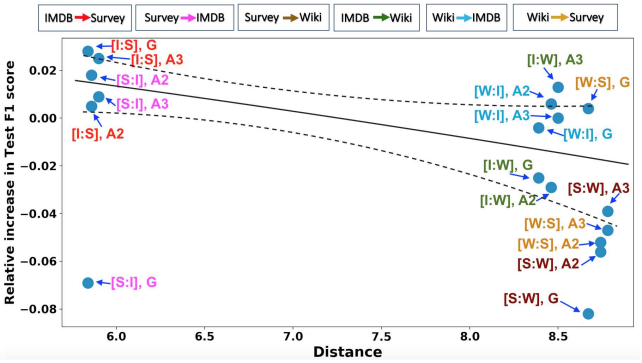


Figure 4: Distance vs. adaptation (S: Survey, W: Wiki, I: IMDB; G: gender inference, A2: 2-bin age inference, A3: 3-bin age inference. Colors represent DITL between data sets.)

data set	Gender		Bin2		Bin3	
	TI	TM	TI	TM	TI	TM
Wiki	0.817±0.026	0.816±0.02	0.829±0.014	0.79±0.033	0.628±0.027	0.633±0.04
IMDB	0.833±0.021	0.856±0.035	0.771±0.008	0.786±0.034	0.59±0.024	0.605±0.04
Survey	0.712 ±0.022	0.751±0.025	0.739±0.024	0.76±0.033	0.629±0.031	0.64±0.036

Table 4: F1 score for the sample data sets. Results are in m±c format, where m=mean and c= 95% confidence interval

distance between two data sets increases, the F1 score improvement brought by transfer learning from one data set to the other becomes smaller. We observe that when the distribution distance between data sets is less than 6, there is usually a gain in F1 score; when the distance is greater than 8.5, in most of the cases, there is a loss; when the distance is in between, we note that there are both gains and losses. The performance shows the correctness of user-level OTDD and also suggests that a model trained on a small data set can still add knowledge to the training of another data set if the two data sets are similar enough and the distribution distance is small. In these cases, combining two small training data sets is a reasonable strategy.

Model	IMDB				Survey		
	Gender	Bin2	Bin3	Bin4	Gender	Bin2	Bin3
RF	0.835	0.717	0.583	0.445	0.68	0.663	0.509
RF_32	0.714	0.582	0.539	0.358	0.574	0.447	0.376
RF_ALL	0.834	0.715	0.565	0.445	0.683	0.671	0.518
BERT	0.812	0.689	0.576	0.435	0.704	0.691	0.517
BERT_32	0.725	0.604	0.315	0.2	0.651	0.577	0.291
BERT_ALL	0.83	0.674	0.555	0.441	0.694	0.661	0.507
Emoji	0.828	0.697	0.561	0.452	0.679	0.721	0.555
Emoji_32	0.853	0.689	0.563	0.461	0.73	0.721	0.552
Emoji_ALL	0.868	0.718	0.602	0.49	0.742	0.734	0.573
Proposed BERT	0.813	0.688	0.558	0.456	0.678	0.708	0.563
Proposed BERT_32	0.845	0.688	0.553	0.465	0.696	0.722	0.555
Proposed BERT_ALL	0.868	0.714	0.582	0.485	0.714	0.731	0.579
Proposed CLIP	0.871	0.754	0.617	0.512	0.712	0.739	0.629
Proposed CLIP_32	0.873	0.733	0.649	0.541	0.745	0.731	0.634
Proposed CLIP_All	0.883	0.761	0.673	0.58	0.756	0.764	0.65

Table 5: Comparison of F1 score for training from scratch without using transfer learning, DITL using 32 samples and then all the data.

4.4 Results Using DITL

To better understand the impact of different components of our transfer learning approach, we begin by showing results for our approach without regularization. We then present results that incorporate the regularization component. For the results in this section, we apply DITL from the larger data set (Wiki) to one of the smaller data sets (IMDB or Survey).

4.4.1 DITL without Regularization. For unregularized DITL, a comparison of F1 scores for different data sets is presented in Table 5. The first row for each model is the original model trained from scratch without transfer learning. The next row shows the results using 32 samples per class from the target training set. The final row shows the results using all the data.[10]

IMDB Analysis: From Table 5, we see that DITL from classic models or simple neural network models does not make enough of a difference when using all the data to fine-tune, and the performance is downgraded by a large percentage when using limited samples. This indicates that a simple model is not able to effectively store and transfer the knowledge needed for the demographic inference task. For the BERT emoji model, we can see that when only using 32 samples, the performance is improved by more than 2% over the one trained from scratch for gender. For age classification, DITL fine-tuned on 32 users achieves comparable results to the original model without transfer learning. While using all the data, we observe that

[10]Due to space limitations, we only show the best performing models.

Model	Pearson	Spearman
Vanilla BERT	0.61	0.587
CLIP	0.686	0.679

Table 6: Pearson and Spearman correlation for BERT and CLIP

Model	Wiki	IMDB	Survey
Latest Post	0.665	0.608	0.551
Random Post	0.663	0.604	0.554
All posts	0.898	0.871	0.712

Table 7: Performance of proposed model using CLIP for gender using the latest post, a random post and all posts

it is better than the original BERT emoji model by 4%, 2.1%, 4.1%, 3.8% for gender, 2-bin, 3-bin and 4-bin, respectively.

For the proposed models, we find that the proposed BERT model has a similar performance as BERT emoji. For the CLIP model, except for the 2-bin case for age, DITL with 32 user samples performs similar to or better than the original model for all the other prediction tasks, with a marginal difference for gender, and an improvement of approximately 3% for multi-bin age classification. When using all the training data, DITL achieves better results than the model without using transfer learning, improvements range from 0.7% to 6.8%. These results indicate that although the magnitude of improvement differs for BERT and CLIP, with a more complex model, using DITL can improve the F1 scores in general. Furthermore, even with a reduced training data set, the performance is still competitive to having the complete training data set.

Survey Data Analysis: For the Survey data set, we observe similar results as the IMDB data set. Both RF and vanilla BERT fail to transfer the knowledge. For BERT emoji, when using 32 samples, the results are comparable to the original model for age, and there is an improvement of approximately 5% for gender. When using all the target training set, we note that both age and gender increase in F1 score over the model trained from scratch, by 6.3%, 1.3%, and 1.8% for gender, age with 2 bins and 3 bins, respectively. For the proposed models, we see a similar performance. Overall, DITL using CLIP achieves a higher result than the original model, by 2.1% to 4.4%. Finally, similar to DIDC, using our proposed model with CLIP performs better than with BERT.

4.4.2 The Impact of Regularization. We now consider the impact of regularization within our proposed model. Following the work of Zhou et al. [36], we use the train-validation difference and train-test difference to show the stability of a model. A larger difference in F1 score implies more overfitting: performing well on the training set and not as well on the test/validation set.[11] From Figure 5, we compare the training dynamics of unregularized and regularized transfer learning with $\alpha = 0.1$ and $\alpha = 1$.[12] As can be seen, under a large regularization parameter $\alpha = 1$, our method achieves smaller differences between the training data F1 and the validation data F1 than the unregularized DITL. Our method also achieves smaller differences between the training F1 and the test F1. Under a smaller regularization parameter $\alpha = 0.1$, differences are lower in some cases compared to unregularized DITL.

Table 8 shows the comparison of relative increase in F1 score for gender and age for unregularized DITL and the regularized DITL as we increase the regularization parameter α from 0.1 to 1. As we increase α, the F1 score generally increases, although the scores are within the same range. This is because α imposes a regularization effect which helps reduce overfitting. We see that in most cases, the regularized model has a marginal improvement over the unregulated model. However, if α becomes too large, the F1 score drops because the regularization effect is too strong.

4.5 Discussion of CLIP vs BERT

Our intuition is that CLIP is able to recognize more similarities and dissimilarities for a post than BERT. To demonstrate this, we compare the performance of BERT and CLIP in terms of their semantic textual similarity. Specifically, we use the public data set, SICK-Relatedness [15] and show the Pearson correlation and Spearman correlation of the two models. The SICK data set consists of 10,000 English sentence pairs with the sentence relatedness score (on a 5-point rating scale). Table 6 presents the Pearson correlation score and Spearman correlation score. We can see that CLIP is higher on both metrics than BERT, suggesting that CLIP's ability to distinguish features of users' posts might be one of the reasons why it performs better.

4.6 Comparison of User-level and Post-level Inference

To further show the difference between post-level and user-level inference, we compare the performance when using the latest tweet, a random tweet and all tweets. Specifically, for the post-level inference, we map the tweet into the embedding space using CLIP. Then, we use the same multilayer perceptron (MLP) model presented in Liu et. al [14]. Table 7 shows the F1 scores. We see that user-level inference works much better than post-level inference for our task. The results suggest that post-level inference is insufficient for this task and that user-level inference is necessary.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose two approaches for dealing with the overfitting problem caused by small data sets when training deep learning models for gender and age. We show that data combination (DIDC) is a good option when we have data sets that are similar in both size and distribution. Using user-level OTDD, we are able to quantify the performance of the combination. We also empirically demonstrate that supervised transfer learning works particularly well on user-level demographic inference, in spite of low or high similarity. We also show that having an advanced neural model with regularization is better for user-level transfer learning in this constrained environment. Finally, using a combined text and image embedding space works well for this task. CLIP has superior performance compared with classic models and BERT, indicating that for Twitter where images are popular, it is beneficial to integrate visual

[11] We only show the results for the proposed model using CLIP as it has the best F1 scores, but the regularized DITL can also be applied to models that use BERT.
[12] We use 10 epochs for these comparison results.

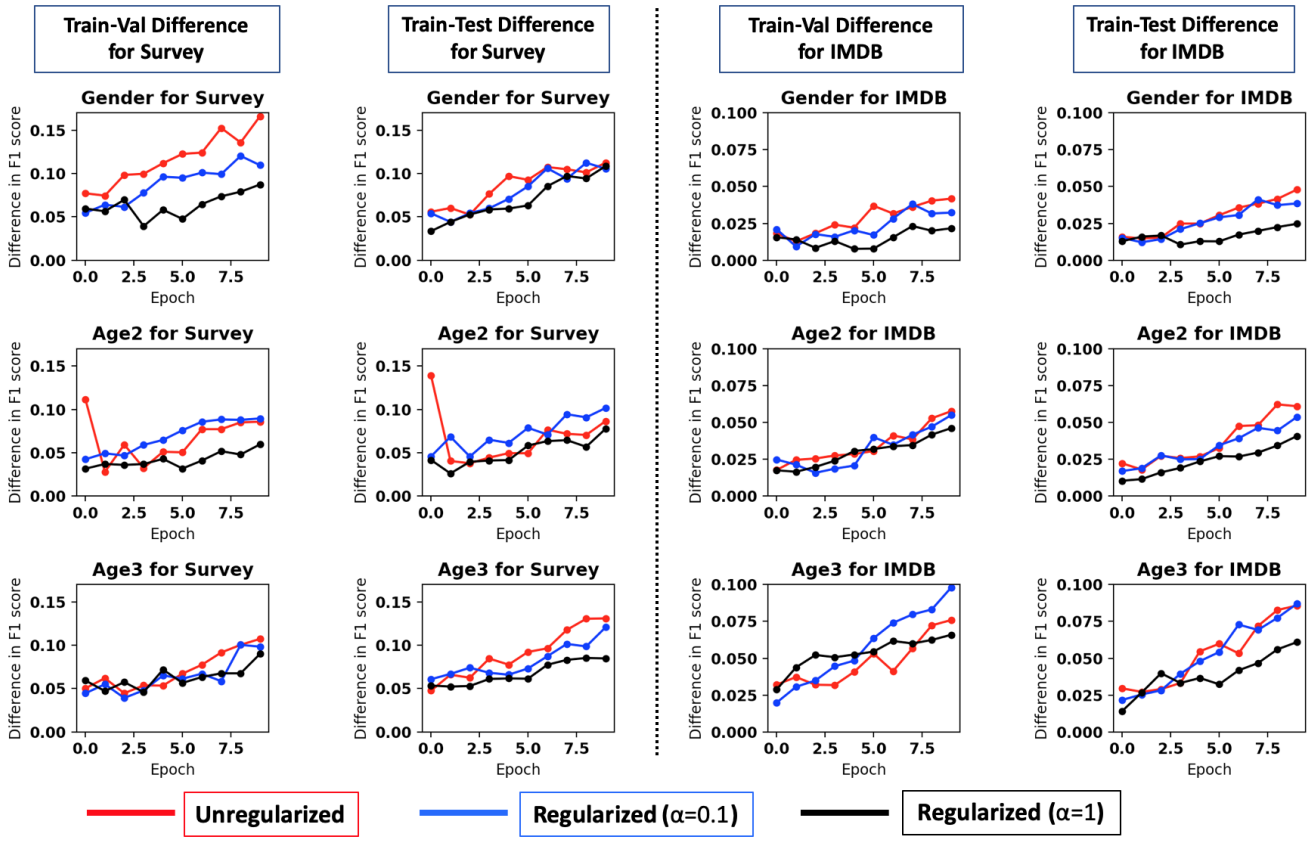


Figure 5: Training dynamics of DITL. The red line denotes unregularized DITL, blue is DITL with α = 0.1 and black is DITL with α = 1. The left 2 columns are for the Survey data and the right 2 are for IMDB. From the top to bottom, each row corresponds to gender, 2-bin age and 3-bin age. For each subfigure, the X axis is the number of epochs and the Y axis is the difference in F1 score.

α	IMDB				Survey		
	Gender	Bin2	Bin3	Bin4	Gender	Bin2	Bin3
N/A	0.884 ± 0.009	0.762 ± 0.01	0.662 ± 0.021	0.546 ± 0.013	0.756 ± 0.045	0.764 ± 0.017	0.65 ± 0.035
0.1	0.885 ± 0.007	0.769 ± 0.013	0.652 ± 0.016	**0.568** ± 0.028	**0.765** ± 0.019	0.756 ± 0.036	**0.664** ± 0.035
0.2	0.888 ± 0.01	**0.771** ± 0.009	**0.663** ± 0.018	0.561 ± 0.021	0.724 ± 0.023	0.764 ± 0.015	0.637 ± 0.029
0.5	**0.892** ± 0.009	0.763 ± 0.012	0.659 ± 0.012	0.554 ± 0.009	0.754 ± 0.03	**0.766** ± 0.028	0.644 ± 0.014
1	**0.892** ± 0.01	0.766 ± 0.01	0.652 ± 0.014	0.555 ± 0.034	0.756 ± 0.03	0.743 ± 0.025	0.655 ± 0.025

Table 8: F1 score with 0.95 confidence interval for unregularized DITL (α is N/A) and regularized DITL with different α values

information. Future work includes considering other demographics with a similar resource constraint.

6 ETHICAL CONSIDERATIONS

We acknowledge that demographic prediction has ethical implications. While automated models could provide valuable information on understanding people's opinions, errors occur that may lead to possible equity and justice related consequences. We also believe that privacy expectations should not be compromised. For this reason we use data sets that either have an expectation of being public (Wiki and IMDB) or ones we obtain consent to use for research purposes (Survey). We also choose to run all of our experiments on Twitter, where users do not typically have an expectation of privacy and where they readily share information that would let other users infer their gender and age, e.g. photos.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation awards #1934925 and #1934494, the National Collaborative on Gun Violence Research (NCGVR) and the Massive Data Institute (MDI) at Georgetown University.

REFERENCES

[1] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. C Fowlkes, S. Soatto, and P. Perona. 2019. Task2vec: Task embedding for meta-learning. In *ICCV*.

[2] D. Alvarez-Melis and N. Fusi. 2020. Geometric dataset distances via optimal transport. *NIPS*.

[3] S. Banerjee, C. Akkaya, F. Perez-Sorrosal, and K. Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *ACL*.

[4] C. Budak, S. Soroka, L. Singh, M. Bailey, L. Bode, N. Chawla, P. Davis-Kean, M. De Choudhury, R. De Veaux, U. Hahn, et al. 2021. Modeling Considerations for Quantitative Social Science Research Using Social Media Data. (2021).

[5] X. Chen, Y. Wang, E. Agichtein, and F. Wang. 2015. A comparative study of demographic attribute inference in twitter. In *ICWSM*.

[6] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

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

[9] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[10] A. Klein, A. Magge, and G. Gonzalez-Hernandez. 2022. ReportAGE: Automatically extracting the exact age of Twitter users based on self-reports in tweets. *PloS one* (2022).

[11] G. Lemaî, F. Nogueira, and C. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* (2017).

[12] D. Levinson. 1986. A conception of adult development. *American psychologist* (1986).

[13] Y. Liu and L. Singh. 2021. Age Inference Using A Hierarchical Attention Neural Network. In *CIKM*.

[14] Y. Liu, L. Singh, and Z. Mneimneh. 2021. A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users. In *DeLTA*.

[15] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*.

[16] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *ACL*.

[17] Z. Mneimneh, J. Pasek, L. Singh, R. Best, L. Bode, E. Bruch, C. Budak, P. Davis-Kean, K. Donato, N. Ellison, et al. 2021. Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data. (2021).

[18] L. Mou, Z. Meng, R. Yan, G. Li, Y. Xu, L. Zhang, and Z. Jin. 2016. How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111* (2016).

[19] A. Ng. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *ICML*.

[20] D. Nguyen, R. Gravel, and T. Trieschnigg, D.and Meder. 2013. " How old do you think I am?" A study of language and age in Twitter. In *ICWSM*.

[21] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[23] D. Rao, D. Yarowsky, et al. 2010. Detecting latent user properties in social media. In *NIPS MLSN Workshop*.

[24] S. Sakaki, Y. Miura, X. Ma, K. Hattori, and T. Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Workshop on Vision and Language*.

[25] T. Semwal, P. Yenigalla, G. Mathur, and S. Nair. 2018. A practitioners' guide to transfer learning for text classification using convolutional neural networks. In *SDM*.

[26] J. Shang, M. Sun, and K. Collins-Thompson. 2018. Demographic inference via knowledge transfer in cross-domain recommender systems. In *ICDM*.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* (2014).

[28] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

[30] P. Vijayaraghavan, S. Vosoughi, and D. Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *ACL*.

[31] Z. Wang, S. Hale, D. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *WWW*.

[32] J. Wei and K. Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196* (2019).

[33] Y. Xu, A. Noy, M. Lin, Q. Qian, H. Li, and R. Jin. 2020. WeMix: How to Better Utilize Data Augmentation. *arXiv preprint arXiv:2010.01267* (2020).

[34] M. Yurochkin, S. Claici, E. Chien, F. Mirzazadeh, and J. Solomon. 2019. Hierarchical optimal transport for document representation. *NeurIPS* (2019).

[35] Y. Zheng. 2015. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data* (2015).

[36] M. Zhou, Z. Li, and P. Xie. 2021. Self-supervised Regularization for Text Classification. *TACL* (2021).