Neural User Factor Adaptation for Text Classification: Learning to Generalize Across Author Demographics

Xiaolei Huang and **Michael J. Paul**
Information Science
University of Colorado
Boulder, CO 80309, USA
{xiaolei.huang,mpaul}@colorado.edu

Abstract

Language usage varies across different demographic factors, such as gender, age, and geographic location. However, most existing document classification methods ignore demographic variability. In this study, we examine empirically how text data can vary across four demographic factors: gender, age, country, and region. We propose a multitask neural model to account for demographic variations via adversarial training. In experiments on four English-language social media datasets, we find that classification performance improves when adapting for user factors.

1 Introduction

Different demographic groups can show substantial linguistic variations, especially in online data (Goel et al., 2016; Johannsen et al., 2015). These variations can affect natural language processing models such as sentiment classifiers. For example, researchers found that women were more likely to use the word *weakness* in a positive way, while men were more likely to use the word in a negative expression (Volkova et al., 2013).

Models for text classification, the automatic categorization of documents into categories, typically ignore attributes about the authors of the text. With the growing amount of text generated by users online, whose personal characteristics are highly variable, there has been increased attention to how user demographics are associated with the text they write. Promising recent studies have shown that incorporating demographic factors can improve text classification (Volkova et al., 2013; Hovy, 2015; Yang and Eisenstein, 2017; Li et al., 2018). Lynn et al. (2017) refer to this idea as *user factor adaptation* and proposed to treat this as a domain adaptation problem in which demographic attributes constitute different domains. We extend this line of work in a number of ways:

- We assemble and publish new datasets containing four demographic factors: gender, age, country, and US region. The demographic attributes are carefully inferred from profile information that is separate from the text data.
- We experiment with neural domain adaptation models (Ganin et al., 2016), which may provide better performance than the simpler models used in prior work on user factor adaptation. We also propose a new model using a multitask framework with adversarial training.
- Our approach requires demographic attributes at training time but not at test time: we learn a single representation to be invariant to demographic changes. This approach thus requires fewer resources than prior work.

In this study, we treat adapting across the demographic factors as a domain work problem, in which we consider each demographic factor as a domain. We focus on four different demographic factors (gender, age, country, region) in four English-language social media datasets (Twitter, Amazon reviews, Yelp hotel reviews, and Yelp restaurant reviews), which contain text authored by a diversity of demographic groups.

We first conduct an exploratory analysis of how different demographic variables are associated with documents and document labels (Section 2). We then describe a neural model for the task of document classification that adapts to demographic factors using a multitask learning framework (Section 3). Specifically, the model is trained to predict the values of the demographic attributes from the text in addition to predicting the document label. Experiments on four social media datasets show that user factor adaptation is important for document classification, and that the proposed model works well compared to alternative domain adaptation approaches (Section 4).

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 136–146
Minneapolis, June 6–7, 2019.

2 Exploratory Analysis of User Factors

We begin with an empirical analysis of how text is related to various demographic attributes of its authors. We first present a description of the demographic attributes. We then conduct qualitative analyses of demographic variations within the collected data on three cascading levels: document, topic and word. The goal is to get a sense of the extent to which language data varies across different user factors and how these factors might interact with document classification. This will motivate our adaptation methods later and provide concrete examples of the user factors that we have in mind.

2.1 Data

We experiment with four corpora from three social media sources:

- **Twitter:** Tweets were labeled with whether they indicate that the user received an influenza vaccination (i.e., a flu shot) (Huang et al., 2017), used in a recent NLP shared task (Weissenbacher et al., 2018).
- **Amazon:** Music reviews from Amazon labeled with sentiment.
- **Hotel:** Hotel reviews from Yelp labeled with sentiment.
- **Restaurant:** Restaurant reviews from Yelp labeled with sentiment.

The latter three datasets were collected for this study. All documents are given binary labels. For the Amazon and Yelp data, we encode reviews with a score >3 (out of 5) as positive and ≤ 3 as negative. For the Yelp data, we removed reviews that had fewer than ten tokens or a helpfulness/usefulness score of zero.

2.1.1 User Attribute Inference

Previous work on user factor adaptation considered the factors of gender, age, and personality (Lynn et al., 2017). We similarly consider gender and age, and instead of personality, we consider a new factor of geographic location. For location, we consider two granularities as different factors, country and region.

These factors must be extracted from the data. One of our goals is to infer these factors in a way that is completely independent of the text used for classification. This is in contrast with the approach used by Lynn et al. (2017), who inferred the attributes from the text of the users, which could arguably confound the interpretation of the results, as domains are defined using the same information available to the classifier. Thus, we used only information from user profiles to obtain their demographic attributes.

Gender and Age. We inferred user gender and age through the user's profile image using the Microsoft Facial Recognition API.[1] Recent comparisons of different commercial face APIs have found the Microsoft API to be the most accurate (Jung et al., 2018) and least biased (Buolamwini and Gebru, 2018). We filtered out users that are inferred to be younger than 12 years old. If multiple faces are in an image, we used the first result from the API. Gender is encoded with two values, male and female. For simplicity, we also binarized the age values ($\leq$30 and $>$30).

Country and Region. We define two factors based on the location of the user. For the Twitter data, we inferred the location of each user with the Carmen geolocation system (Dredze et al., 2013), which resolves the user's location string in their profile to a structured location. Because this comes from the user profile, it is generally taken to be the "home" location of the user. For Amazon and Yelp, we collected user locations listed in their profiles, then used pattern matching and manual whitelisting to resolve the strings to specific locations (city, state, country). To construct user factors from location data, we first created a binary country variable to indicate if the user's country is the United States (US, the most common country in the data) or not. Among US users, we resolved the location to a region. We follow the US Census Bureau's regional divisions (Bureau, 2012) to categorize the users into four regional categories: Northeast (NE), Midwest (MW), South (S) and West (W). We labeled Washington D.C. as northeast in this study; we excluded other territories of the US, such as Puerto Rico and U.S. Virgin Islands, since these locations do not contain much data and do not map well to the four regions.

Accuracy of Inference Attributes inferred with these tools will not be perfectly accurate. Although such inaccuracies could lead to suboptimal

[1] https://azure.microsoft.com/en-us/services/cognitive-services/face/

training, this does not affect our classifier evaluation, since we do not use demographic labels at test time. Nonetheless, we provide a rough estimate of the accuracy of the attributes extracted from faces. We randomly sampled 100 users across our datasets. Two annotators reviewed each image and guessed the gender and age of the user (using our binary categories) based on the profile image. A third annotator chose the final label when the first two disagreed (annotators disagreed on gender in 2% of photos and age in 15% of photos). Our final annotations agreed with the Face API's gender estimates 88% of the time across the four datasets (ranging from 84% to 100%), and age estimates 68% of the time across the four datasets (ranging from 56% to 92%).

2.1.2 Data Summary

We show the data statistics along with the full demographic distributions in the Table 1. While our study does not require a representative sample from the data sources, since our primary goal is to evaluate whether we can adapt models to different demographics, we observe some notable differences between the demographics of our collection and the known demographics of the sources. Namely, the percentage of female users is much higher in our data than among Twitter users (Tien, 2018) and Yelp users (Yelp, 2018) as estimated from surveys. This discrepancy could stem from our process of sampling only users who had profile images available for demographic inference, since not all users provide profile photos, and those who do may skew toward certain demographic groups (Rose et al., 2012).

2.1.3 Privacy Considerations

While our data collection includes only public data, due to the potential sensitivity of user profile information, we stored only data necessary for this study. Therefore, we anonymized the personal information and deleted user images after retrieving the demographic attributes from the Microsoft API. We only include aggregated information in this paper and do not publish any private information associated with individuals including example reviews. The dataset that we share will include our model inferences but not the original image data; instead, the dataset will provide instructions on how the data was collected in enough detail that the approach can be replicated.

2.2 Are User Factors Encoded in Text?

It is known that the user factors we consider are associated with variability in language, including in online content (Hovy, 2015). For example, age affects linguistic style (Wagner, 2012), and language styles are highly associated with the gender of online users (Hovy and Purschke, 2018). Dialectical differences also cause language variation by location; for example, "dese" (these) is more common among social media users from the Southern US than other regions of the US (Goel et al., 2016).

Our goal in this section is to test whether these variations hold in our particular datasets, how strong the effects are, and which of our four factors are most associated with language. We do this in two ways, first by measuring predictability of factors from text, and second by qualitatively examining topic differences across user groups.

2.2.1 User Factor Prediction

We explore how accurately the text documents can predict user demographic factors. We do this by training classifiers to predict each factor. We first downsample without replacement to balance the data for each category. We shuffle and split the data into training (70%) and test (30%) sets. We then build logistic regression classifiers using TF-IDF-weighted 1-, 2-, and 3-grams as features. We use *scikit-learn* (Pedregosa et al., 2011) to implement the classifiers and accuracy scores to measure the predictability. We show the absolute improvements of scores in Table 2.

The results show that user factors are encoded in text well enough to be predicted significantly. Twitter data shows the best predictability towards age, and the two Yelp datasets show strong classification results for both gender and country. We also observe that as the data size increases, the predictability of language usage towards demographic factors also increases. These observations suggest a connection between language style and user demographic factors in large corpora.

2.2.2 Topic Analysis

We additionally examine how the distribution of text content varies across demographic groups. To characterize the content, we represent the text with a topic model. We trained a Latent Dirichlet Allocation (Blei et al., 2003) model with 10 topics using GenSim (Řehůřek and Sojka, 2010) with default parameters. After training the topic model, each document d is associated with a probability

	# Docs	# Users	Gender		Age		Country		Region			
			F	M	≤30	>30	US	¬US	NE	MW	S	W
Twitter	9.8K	9.8K	.575	.425	.572	.428	.772	.228	.104	.120	.145	.631
Amazon	40.4K	34.3K	.333	.667	.245	.755	.900	.100	.097	.096	.132	.675
Hotel	169K	119K	.576	.424	.450	.550	.956	.044	.297	.166	.271	.266
Restaurant	713K	811K	.547	.453	.451	.549	.892	.108	.305	.181	.302	.212

Table 1: Dataset statistics including user demographic distributions for four user factors.

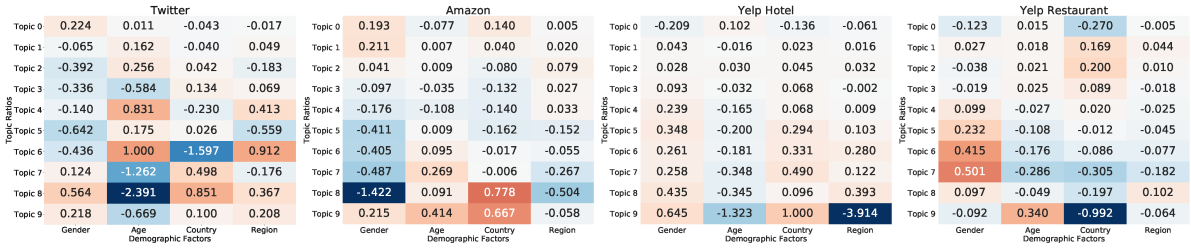

Figure 1: Topic distribution log ratios. A value of 0 means that demographic groups use that topic in equal amounts, while values away from 0 mean that the topic is discussed more by one demographic group than the other group(s) in that factor.

	Gender	Age	Country	Region
Twitter	+9.6	+15.3	+9.0	+3.3
Amazon	+15.2	+12.2	+18.0	+13.0
Hotel	+17.2	+10.9	+25.4	+11.6
Restaurant	+19.0	+13.2	+32.8	+17.5

Table 2: Predictability of user factors from language data. We show the absolute percentage improvements in accuracy over majority-class baselines. For example, the majority-class baselines of accuracy scores are either .500 for the binary prediction or .250 for the region prediction.

distribution over the 10 topics. The model learns a multinomial topic distribution $P(Z|D)$ from a Dirichlet prior, where Z refers to each topic and D refers to each document. For each demographic group, we calculate the average topic distribution across the documents from that group. Then within each factor, we calculate the log-ratio of the topic probabilities for each group. For example, for topic k for the gender factor, we calculate $\log_2 \frac{P(Topic{=}k|Gender{=}\text{female})}{P(Topic{=}k|Gender{=}\text{male})}$. The sign of the log-ratio indicates which demographic group is more likely to use the topic. We do this for all factors; for region, we simply binarize the four values for the purpose of this visualization (MW + W vs. NE + S). Results are shown in Figure 1.

The topic model was trained without removing stop words, in case stop word usage varies by group. However, because of this, the topics all look very similar and are hard to interpret, so we do not show the topics themselves. What we instead want to show is the degree to which the prevalence of some topics varies across demographic attributes, which are extracted independently from the text used to train the topic models. We see that while most topics are fairly consistent across demographic groups, most datasets have at least a few topics with large differences.

2.3 Are Document Categories Expressed Differently by Different User Groups?

While text content varies across different user groups, it is a separate question whether those variations will affect document classification. For example, if men and women discuss different topics online, but express sentiment in the same way, then those differences will not affect a sentiment classifier. Prior work has shown that the way people express opinions in online social media does vary by gender, age, geographic location, and political orientation (Hinds and Joinson, 2018); thus, there is reason to believe that concepts like sentiment will be expressed differently by different groups. As a final exploratory experiment, we now consider whether the text features that are predictive of document categories (e.g., positive or negative sentiment) also vary with user factors.

To compare how word expressions vary among the demographic factors, we conduct a word-level feature comparison. For each demographic group, we collect only documents that belong to that group and then calculate the n-gram features

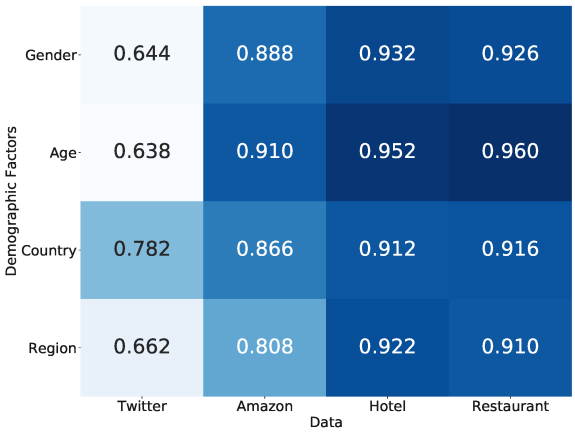

Figure 2: Overlap in most predictive classification features across different demographic groups, calculated for each demographic factor and each dataset. Darker color indicates less variation in word usage across demographic groups.

(same features as in Section 2.2) that are most associated with the document class labels. Using mutual information, we select the top 1,000 features for each attribute. Then within each demographic factor (e.g., gender), we calculate the percentage of top 1,000 features that overlap across the different attribute values in that factor (e.g., male and female). Specifically, if S_0 is the set of top features for one attribute and S_1 is the set of top features for another attribute, the percent overlap is calculated as $|S_0 \cap S_1|/1000$. Results are shown in Figure 2. Lower percentages indicate higher variation in how different groups express the concepts being classified (e.g., sentiment). The Twitter data shows the most variation while the Yelp hotel data shows the least variation.

3 Model

Models for user factor adaptation generally treat this as a problem of *domain adaptation* (Volkova et al., 2013; Lynn et al., 2017). Domain adaptation methods are used to learn models that can be applied to data whose distributions may differ from the training data. Commonly used methods include feature augmentation (Daume III, 2007; Joshi et al., 2013; Huang and Paul, 2018) and structural correspondence learning (Blitzer et al., 2006), while recent approaches rely on domain adversarial training (Ganin et al., 2016; Chen et al., 2016; Liu et al., 2017; Huang et al., 2018). We borrow concepts of domain adaptation to construct a model that is robust to variations across user factors.

In our proposed **Neural User Factor Adaptation (NUFA)** model, we treat each variable of interest (demographic attributes and document class label) as a separate, but jointly modeled, prediction task. The goal is to perform well at predicting document classes, while the demographic attribute tasks are modeled primarily for the purpose of learning characteristics of the demographic groups. Thus, the model aims to learn discriminative features for text classification while learning to be invariant to the linguistic characteristics of the demographic groups. Once trained, this classifier can be applied to test documents without requiring the demographic attributes.

Concretely, we propose the multitask learning framework in Figure 3. The model extracts features from the text for the demographic attribute prediction tasks and the classification task, as well as joint features for all tasks in which features for both demographics and document classes are mapped into the same vector space. Each feature space is constructed with a separate Bidirectional Long Short-Term Memory model (Bi-LSTM) (Hochreiter and Schmidhuber, 1997).

Because language styles vary across groups, as shown in Section 2.2, information from each task could be useful to the other. Thus, our intuition is that while we model the document and demographic predictions as independent tasks, the shared feature space allows the model to transfer knowledge from the demographic tasks to the text classification task and vice versa.

However, we want to keep the feature space such that the features are predictive of document classes in a way that is invariant to demographic shifts. To avoid learning features for the document classifier that are too strongly associated with user factors, we use adversarial training. The result is that the demographic information is encoded primarily in the features used for the demographic classifiers, while learning invariant text features that work *across* different demographic groups for the document classifier.

Domain Sampling and Model Inputs. Our model requires all domains (demographic attributes) to be known during training, but not all attributes are known in our datasets. Instead of explicitly modeling the missing data, we simply sample documents where all user attributes of interest are available. At test time, this limitation does not apply because only the document text is

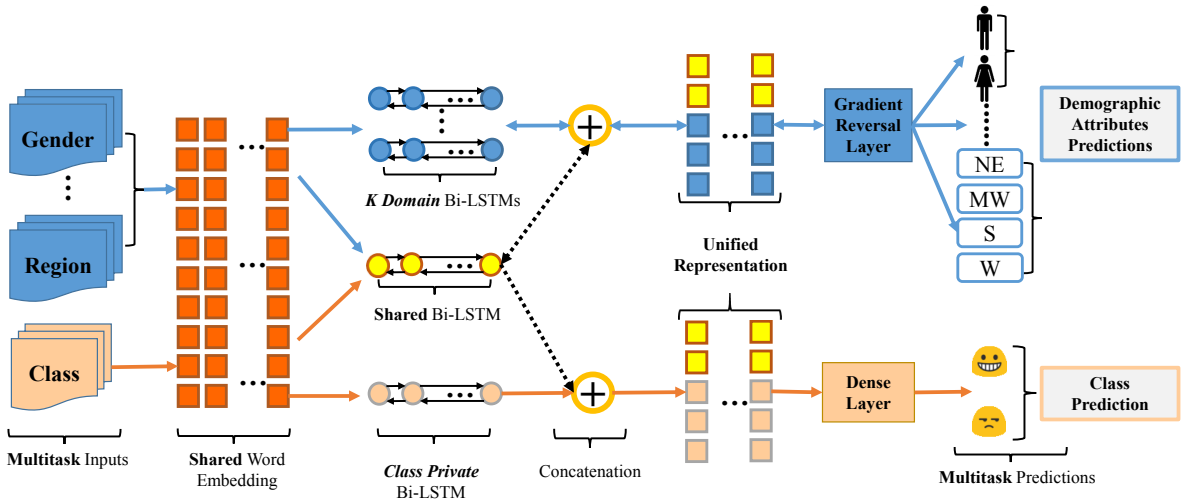

Figure 3: Neural User Factor Adaptation (NUFA) model. NUFA optimizes for two major tasks, demographic prediction (blue blocks and arrows) and text classification (light orange blocks and arrows). During the training phase, documents labeled with demographic information go through the demographic classifier, and documents with class labels go through the document classifier. This helps NUFA learn representations that are useful for classifying documents versus representations that are useful for predicting demographics. At test time, documents are given only to the document classifier, leaving out the demographic classifiers.

required as input to the document classifier.

Shared Embedding Space. We use a common embedding layer for both document and demographic factor predictions. The goal is that the trained embeddings will capture the language variations that are associated with the demographic groups as well as document labels. Parameters are initialized with pre-trained embeddings (Mikolov et al., 2013; Pennington et al., 2014).

K+2 Bi-LSTMs. We combine ideas from two previous works on domain adaptation (Liu et al., 2017; Kim et al., 2017). Kim et al. (2017) proposed $K+1$ Bi-LSTMs, where K is the number of domains, and Liu et al. (2017) proposed to combine shared and independent Bi-LSTMs for each prediction task. In our model, we create one independent Bi-LSTM for each demographic domain (blue), one independent Bi-LSTM for the document classifier (orange), and one shared Bi-LSTM that is used in both the demographic prediction and document classification tasks (yellow). The intuition is to transfer learned information to one and the other through this shared Bi-LSTM while leaving some free spaces for both document label and demographic factors predictions. We then concatenate outputs of the shared LSTM with each task-independent LSTM together. This helps the text classifier capture demographic knowledge.

Demographic Classifier. We adjust the degree to which the demographic classifiers can discriminate between attributes. To find a balance between the invariant knowledge and differences across user demographic factors, we apply domain adversarial training (Ganin et al., 2016) (the blue block indicating the "gradient reversal layer") to each domain prediction task. The predictions use the final concatenated representations, where the prediction is modeled with a softmax function for the region and a binary sigmoid function for the other user demographic factors.

Document Classifier. We feed the concatenated outputs of the document and shared Bi-LSTMs to one layer feed-forward network (the orange block indicating the "dense layer"). Finally, the document classifier outputs a probability via a sigmoid.

Joint Multitask Learning. We use the categorical cross-entropy loss to optimize the $K+1$ prediction tasks jointly. One question is how to assign importance to the multiple tasks. Because our target is document classification, we assign a cost to the domain prediction loss (L_{domain}). Each prediction task has its own weight, α_k. The final loss function is defined as $L = L_{doc} + \sum_{k=1}^{K} \alpha_k L_{domain,k}$. In summary, the proposed model learns and adapts to user demographic factors through three aspects: shared embeddings, shared Bi-LSTMs, and joint optimization.

4 Experiments

We experiment with document classification on our four corpora using various models. Our goal is to test whether models that adapt to user factors can outperform models that do not, and to understand which components of models can facilitate user factor adaptation.

4.1 Data Processing

We replaced hyperlinks, usernames, and hashtags with generic symbols. Documents were lowercased and tokenized using NLTK (Bird and Loper, 2004). The corpora were randomly split into training (80%), development (10%), and test (10%) sets. We train the models on the training set and find the optimal hyperparameters on the development set. We randomly shuffle the training data at the beginning of each training epoch. The evaluation metric is weighted F1 score.

4.2 Baselines: No Adaptation

We compare to three standard classifiers that do not perform adaptation.

N-gram. We extract TF-IDF-weighted features of 1-, 2-, and 3-grams on the corpora, using the most frequent 15K features with the minimum feature frequency as 2. We trained a logistic regression classifier using the `SGDClassifier` implementation in scikit-learn (Pedregosa et al., 2011) using a batch size of 256 and 1,000 iterations.

CNN. We used Keras (Chollet et al., 2015) to implement the Convolutional Neural Network (CNN) classifier described in Kim (2014). To keep consistent, we initialize the embedding weight with pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014). We only keep the 15K most frequent words and replace the rest with an "unk" token. Each document was padded to a length of 50. We keep all parameter settings as described in the paper. We fed 50 documents to the model each batch and trained for 20 epochs.

Bi-LSTM. We build a bi-directional Long Short Term Memory (bi-LSTM) (Hochreiter and Schmidhuber, 1997) classifier. The classifier is initialized with the pre-trained word embeddings, and we initialize training with the same parameters used for the NUFA.

4.3 Adaptation Models

We consider two baseline domain adaptation models that can adapt for user factors, a non-neural method and a neural model. We then provide the training details of our proposed model, NUFA. Finally, we consider two variants of NUFA that ablate components of the model, allowing us to evaluate the contribution of each component.

FEDA. Lynn et al. (2017) used a modification of the "frustratingly easy" domain adaptation (FEDA) method (Daume III, 2007) to adapt for user factors. We use a modification of this method where the four user factors and their values are treated as domains. We first extract domain-specific and general representations as TF-IDF-weighted n-gram (1-, 2, 3-grams) features. We extract the top 15K features for each domain and the general feature set. With this method, the feature set is augmented such that each feature has a domain-specific version of the feature for each domain, as well as a general domain-independent version of the feature. The features values are set to the original feature values for the domain-independent features and the domain-specific features that apply to the document, while domain-specific features for documents that do not belong to that domain are set to 0. For example, using gender as a domain, a training document with a female author would be encoded as $[F_{general}, F_{domain,female}, 0]$, while a document with a male author would be encoded as $[F_{general}, 0, F_{domain,male}]$. Different from prior work with FEDA for user-factor adaptation, at test time we only use the general, domain-independent features; the idea is to learn a generalized feature set that is domain invariant. This is the same approach we used in recent work using FEDA to adapt classifiers to temporal variations (Huang and Paul, 2018).

DANN. We consider the domain adversarial training network (Ganin et al., 2016) (DANN) on the user factor adaptation task. We use Keras to implement the same network and deploy the same pre-trained word embeddings as in NUFA. We then set the domain prediction as the demographic factors prediction and keep the document label prediction as the default. We train the model with 20 epochs with a batch size of 64. Finally, we use the model at the epoch when the model achieves the best result on the development set for

the final model.

NUFA. We initialize the embedding weights by the pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) with 200 dimensional vectors. All LSTMs are fixed outputs as 200-dimension vectors. We set the dropout of LSTM training to 0.2 and the flip gradient value to 0.01 during the adversarial training. The dense layer has 128 neurons with ReLU activation function and dropout of 0.2. User factors and document label predictions are optimized jointly using Adam (Kingma and Ba, 2015) with a learning rate of 0.001 and batch size of 64. We train NUFA for up to 20 epochs and select the best model on the development set. For single-factor adaptation (next section), we set α to 0.1; for multi-factor adaptation, we use a heuristic for setting α described in that section. We implemented NUFA in Keras (Chollet et al., 2015).

NUFA–s. To understand the role of the shared Bi-LSTM in our model, we conduct experiments on NUFA without the shared Bi-LSTM. We follow the same experimental steps as NUFA and denote it as NUFA−s (NUFA minus shared Bi-LSTM).

NUFA–a. To understand the role of the adversarial training in our model, we conduct experiments of the NUFA without adversarial training, denoted as NUFA−a (NUFA minus adversarial).

4.4 Results

4.4.1 Single-Factor Adaptation

We first consider user factor adaptation for each of the four factors individually. Table 3 shows the results. Adaptation methods almost always outperform the non-adaptation baselines; the best adaptation model outperforms the best non-adaptation model by 1.5 to 5.5 points. The improvements indicate that adopting the demographic factors might be beneficial for the classifiers. User factor adaptation thus appears to be important for text classification.

Comparing the adaptation methods, our proposed model (NUFA) is best on three of four datasets. On the Hotel dataset, the n-gram model FEDA is always best; this seems to be a dataset where neural methods perform poorly, since even the n-gram baseline with no adaptation often outperformed the various neural models. Whether a neural model is the best choice depends on the dataset, but among the neural models, NUFA always outperforms DANN. Finally, the full NUFA model most often outperforms the variants without the shared Bi-LSTM (NUFA−s) and without adversarial training (NUFA−a).

	Twitter	Amazon	Hotel	Rest.
No Adaptation				
N-gram	.866	.793	.857	.866
CNN	.879	.776	.825	.846
Bi-LSTM	.869	.776	.842	.875
Adaptation (Gender)				
FEDA	.814	.809	**.865**	.874
DANN	.864	.832	.813	.855
NUFA−s	.880	*.845*	.857	.869
NUFA−a	.874	.842	.852	.868
NUFA	*.886*	.844	.854	**.881**
Adaptation (Age)				
FEDA	.813	.801	**.865**	.873
DANN	.856	.824	.811	.851
NUFA−s	.872	*.843*	.850	.879
NUFA−a	.882	.841	.852	.878
NUFA	*.885*	.839	.857	*.880*
Adaptation (Country)				
FEDA	.826	.768	**.865**	.877
DANN	.868	.828	.827	.855
NUFA−s	.882	*.844*	.854	*.879*
NUFA−a	.880	.838	.855	.877
NUFA	**.896**	.843	.854	*.879*
Adaptation (Region)				
FEDA	.826	.780	*.864*	.869
DANN	.875	.825	.823	.852
NUFA−s	.874	.833	.854	.878
NUFA−a	.882	.838	.854	.875
NUFA	*.893*	**.848**	.853	*.880*

Table 3: Performance (weighted F1) of no adaptation and single user factor adaptation. For each dataset, the best score within each demographic domain is italicized; the best score overall is bolded.

4.4.2 Multi-Factor Adaptation

Finally, we experiment with adapting to all four user factors together. Recall that each domain prediction task in NUFA is weighted by α_k. Initially, we simply used a uniform weighting, $\alpha_k = \alpha/K$, but we find that we can improve performance with non-uniform weighting. Because optimizing the α vector would be expensive, we instead propose a heuristic that weighs the domains based on how much each domain is expected to influence the text. We define $\alpha_k = s_k/(\sum_{k'} s_{k'})$, where s_k

	Twitter	Amazon	Hotel	Rest.
Baseline Adaptation				
FEDA	.806	.778	**.867**	.869
DANN	.880	.828	.830	.858
Proposed Model				
NUFA	.887	.848	.853	.879
NUFA+w	**.901**	**.852**	.855	**.885**

Table 4: Results of adaptation for all four user factors.

is the F1 score of demographic attribute prediction for domain k from Table 2. We denote this method as **NUFA+w**, which refers to this additional weighting process.

Table 4 shows that combining all user factors provides a small gain over single-factor adaptation; the best multi-factor result is higher than the best single-factor result for each dataset. As with single-factor adaptation, FEDA works best for the Hotel datasets, while NUFA+w works best for the other three. Without adding weighting to NUFA, the multi-factor performance is comparable to single-factor performance; thus, task weighting seems to be critical for good performance when combining multiple factors.

5 Related Work

Demographic prediction is a common task in natural language processing. Research has shown that social media text is predictive of demographic variables such as gender (Rao et al., 2010, 2011; Burger et al., 2011; Volkova et al., 2015) and location (Eisenstein et al., 2010; Wing and Baldridge, 2011, 2014). Our work is closely related to these, as our model also predicts demographic variables. However, in our model the goal of demographic prediction is primarily to learn representations that will make the document classifier more robust to demographic variations, rather than the end goal being demographic prediction itself.

Demographic bias has been shown to be encoded in machine learning models. Word embeddings, which are widely used in classification tasks, are prone to learning demographic stereotypes. For example, a study by Bolukbasi et al. (2016) found that the word "programmer" is more similar to "man" than "woman," while "receptionist" is more similar to "woman." To avoid learning biases, researchers have proposed adding demographic constraints (Zhao et al., 2017) or using adversarial training (Elazar and Goldberg, 2018). While our work is not focused specifically on reducing bias, our goals are related to it in that our models are meant to learn document classifiers that are invariant to author demographics.

6 Conclusion

We have explored the issue of author demographics in relation to document classification, showing that demographics are encoded in language, and the most predictive features for document classification vary by demographics. We showed that various domain adaptation methods can be used to build classifiers that are more robust to demographics, combined in a neural model that outperformed prior approaches. Our datasets, which contain various attributes including those inferred through facial recognition, could be useful in other research (Section 5). We publish our datasets[2] and source code.[3]

7 Acknowledgements

The authors thank the anonymous reviews for their insightful comments and suggestions. The authors thank Zijiao Yang for helping evaluate inference accuracy of the Microsoft Face API. This work was supported in part by the National Science Foundation under award number IIS-1657338.

References

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

[2] http://cmci.colorado.edu/~mpaul/files/starsem2019_demographics.data.zip

[3] https://github.com/xiaoleihuang/NUFA

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

United States Census Bureau. 2012. 2010 geographic terms and concepts - census divisions and census regions.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. Springer.

Joanne Hinds and Adam N Joinson. 2018. What demographic attributes do our digital footprints reveal? a systematic review. *PloS one*, 13(11).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701.

Xiaolei Huang and Michael J Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 694–699.

Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*, pages 542–546.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112.

Mahesh Joshi, Mark Dredze, William W Cohen, and Carolyn P Rosé. 2013. Whats in a domain? multi-domain learning for multi-attribute data. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–690.

Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard Jim Jansen. 2018. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Twelfth International AAAI Conference on Web and Social Media*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Domain attention with an ensemble of experts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 643–653.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 25–30.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1–10.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. 2017. Human centered nlp with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *International Conference on Weblogs and Social Media (ICWSM)*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Workshop on Search and Mining User-generated Contents*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Jessica Rose, Susan Mackey-Kallis, Len Shyles, Kelly Barry, Danielle Biagini, Colleen Hart, and Lauren Jack. 2012. Face it: The impact of gender on social media images. *Communication Quarterly*, 60(5):588–607.

Shannon Tien. 2018. Top twitter demographics that matter to social media marketers in 2018.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI Conference on Artificial Intelligence (AAAI)*, Austin, TX.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Suzanne Evans Wagner. 2012. Age grading in sociolinguistic theory. *Language and Linguistics Compass*, 6(6):371–382.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 13–16. Association for Computational Linguistics.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348.

Benjamin P Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Association for Computational Linguistics (ACL)*.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association of Computational Linguistics*, 5(1):295–307.

Yelp. 2018. An introduction to yelp metrics as of september 30, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.