Residualized Factor Adaptation for Community Social Media Prediction Tasks

Mohammadzaman Zamani,[1] **H. Andrew Schwartz,**[1] **Veronica E. Lynn,**[1]
Salvatore Giorgi,[2] and **Niranjan Balasubramanian**[1]
[1] Computer Science Department, Stony Brook University
[2] Department of Psychology, University of Pennsylvania
mzamani@cs.stonybrook.edu

Abstract

Predictive models over social media language have shown promise in capturing community outcomes, but approaches thus far largely neglect the socio-demographic context (e.g. age, education rates, race) of the community from which the language originates. For example, it may be inaccurate to assume people in Mobile, Alabama, where the population is relatively older, will use words the same way as those from San Francisco, where the median age is younger with a higher rate of college education. In this paper, we present *residualized factor adaptation*, a novel approach to community prediction tasks which both (a) effectively integrates community attributes, as well as (b) adapts linguistic features to community attributes (factors). We use eleven demographic and socioeconomic attributes, and evaluate our approach over five different community-level predictive tasks, spanning health (heart disease mortality, percent fair/poor health), psychology (life satisfaction), and economics (percent housing price increase, foreclosure rate). Our evaluation shows that *residualized factor adaptation* significantly improves 4 out of 5 community-level outcome predictions over prior state-of-the-art for incorporating socio-demographic contexts.

1 Introduction

Adapting to human factors has been shown to benefit NLP tasks, especially in tasks that involve predictions over individual social media posts (e.g., sentiment (Hovy, 2015), sarcasm, and stance detection (Lynn et al., 2017)). The main idea behind these approaches is that knowing who wrote a piece of text can help models better understand how to process it. This paper develops methods that apply this idea to community-level prediction tasks, which require making decisions over posts from a community of users. Many community-level outcomes and community-wide language are linked to socio-demographic factors (age, gender, race, education, income levels) with many social scientific studies supporting their predictive value (Cohen et al., 2003), and should therefore affect how a model treats social media-based language features. For example, a high prevalence of the word "bike" in San Francisco, CA might be a signal that exercise is common in the area, while its high prevalence in Mobile, Alabama might indicate greater interest in motor bikes. We present a method for building language-based predictive models which integrate in and adapt to attributes of the communities generating the language.

This work aims to unify two different approaches developed for adapting to human factors and use them for incorporating community attributes in community-level prediction tasks: (1) *residualized controls:* whereby a model is trained in two steps: first over the factors/controls and then fitting the language to the residuals of the control model (Zamani and Schwartz, 2017), and (2) *user-factor adaptation:* whereby linguistic features are adapted, or treated differently, based on the continuous-valued factors of the authors of the features (Lynn et al., 2017).

Combining factor adaptation (FA) and residualized control (RC) into RFA is a non-trivial task. The intent behind both methods are quite different: whereas RC attempts to address the inherent heterogeneity between robust control variables and noisy linguistic variables, FA enables a model to treat linguistic features differently depending on the factors. From a statistical learning perspective, RC separates inference over controls from inference over language (model level integration), while FA brings controls and language together and makes the inference as one single step (data level integration). Additionally, FA has stricter bounds in the number of factors it can accommodate because each new factor has a multiplicative

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3560–3569
Brussels, Belgium, October 31 - November 4, 2018.

effect on the number of learned parameters. On the other hand, each new factor for RC typically only adds one new parameter. Here, we endeavour to develop RFA such that it achieves the benefits of both approaches with little lost to the limitations. RFA inherits the challenges of the FA method with feature explosion. We address this through a systematic exploration of both feature and factor selection.

The main contributions of this work include: (1) the introduction of residualized factor adaptation which effectively combines extra-linguistic and language features, (2) the first empirical evaluation of applying factor adaptation for community-level prediction tasks, (3) analysis of the impact of the size of factors and factor selection in adaptation, and (4) state-of-the-art accuracies for each of the five tasks for which we evaluate RFA.

2 Background

Social media provides easy access to a vast amount of language written by a diverse group of users, making it an increasingly popular resource for measuring community health, psychology, and economics (Coppersmith et al., 2015; Eichstaedt et al., 2015; Weeg et al., 2015; Mowery et al., 2016; Haimson and Hayes, 2017). Coppersmith et al. (2015), for instance, examine trends in language use among Twitter users who self-reported one of ten mental health diagnoses. Eichstaedt et al. (2015) and Weeg et al. (2015) use Twitter to predict the prevalence rates of various health outcomes, such as heart disease mortality and depression, at the county level. Haimson and Hayes (2017) tracked changes in the emotional well-being of transgender communities on Tumblr between 2015 and 2016.

Socio-demographics are often correlated with health outcomes (such as age and heart disease), which is why such variables are often used as controls during analysis (Coppersmith et al., 2015; Dos Reis and Culotta, 2015; Eichstaedt et al., 2015; Weeg et al., 2015). Because of their predictive power, socio-demographics and other extra-linguistic information can additionally be leveraged when building the model itself.

However, a central challenge in integrating community attributes is that they have very different properties than linguistic features and can be lost, in essence, like a needle in a haystack. For example, linguistic features like n-grams are high dimensional, with each dimension having high covariance with other dimensions and likely very little relationship with the outcome. On the other hand community features may be measured more robustly and are relatively low dimensional, often obtained through well-defined measurements. Not surprisingly, Zamani and Schwartz (2017) showed a naive combination that simply concatenates these two sets of features risks losing the effective extra-linguistic features in a sea of weak linguistic features. They go on to show a *residualized control* approach achieves significantly greater accuracy at economic prediction by first learning a model using extra-linguistic features (i.e. controls or community factors) and then train a language model on top of the residual error of the previous model.

It is possible that even when extra-linguistic features are not directly beneficial for prediction, they can still affect people's language. Other related works consider how the meaning of language changes depending on who states it. For instance, when an NLP PhD student says the word 'paper' he/she usually means something different than when a 5^{th} grade student uses the same word (i.e. 'research paper' versus 'piece of paper'). Hu et al. (2017) noted the same words can have different meanings if different people say them. This idea of contextualizing language with extra-linguistic information has been the basis for multiple models: Hovy (2015) learn age- and gender-specific word embeddings, leading to significant improvements for three text classification tasks. Volkova et al. (2013) found that using gender-specific features lead to improvements in sentiment analysis over a gender-agnostic model. Most recently, Lynn et al. (2017) proposed a domain adaptation-inspired method for composing user-level, extra-linguistic information with message-level features, leading to improvements for multiple text classification tasks; we build off of this approach and that of Zamani and Schwartz (2017) in this paper.

While Lynn et al. (2017) injected user-level info into message-level tasks, we are investigating whether same-level adaptation techniques are similarly useful.

We also try to find the circumstances under which each of the adaptation and residualized control approaches are more powerful, and we take on the non-trivial task of exploiting concepts from both the adaptation and the residualized control techniques at the same time, finding that they add even more power when combined with one another.

3 Method

We describe *residualized factor adaptation* (RFA), an approach to text-based prediction utilizing extra-linguistic factors (also called controls – often demographic or socioeconomic information). The key challenge for RFA lies in effectively combining two different types of features. The language-based features, extracted from the tweets, are numerous but are only weak indicators of the outcomes. The socioeconomic and demographic features, on the other hand, are strong indicators but fewer in number. Naively combining both sets of features ignores this crucial difference in their predictive abilities, potentially resulting in important features getting drowned out.

We first describe two methods that effectively combine extra-linguistic factors at two different levels: 1) *Residualized control* is a model-level combination method which builds different models for each type of feature, then combines the results of these models to make the final outcome prediction. 2) *Factor adaptation* is a feature-level combination method that composes the two feature sets with one another to produce a transformed feature space over which a single model may be built. Finally, we present our combined method of *Residualized Factor Adaptation* which takes advantage of both concepts without exploding model parameters.

3.1 Residualized Control Prediction

Language-based features and community-level attributes are qualitatively different modalities. The extra-linguistic variables, while few in number, are mostly unbiased and follow a normal distribution, which can be used to build a strong outcome predictor. However, without special treatment, the signal in extra-linguistic variables can be overwhelmed when combined with a large number of language-based features.

The residualized control approach (Zamani and Schwartz, 2017) avoids this issue by building two models. The first is a prediction model built over the extra-linguistic variables (or *controls*) alone. The error, or *residuals*, produced by this first model represents the information that was unable to be predicted using the extra-linguistic variables alone. The language-based features are therefore brought in to improve upon the initial predictions by using the residuals as training labels for a model based on the linguistic features. In this way, the language-based features are able to account for additional information not captured by the initial extra-linguistic feature-only model. At test time, each instance is fed to both prediction models, and the final outcome is given as a sum of the predictions from both models — the outcome predicted by the extra-linguistic model adjusted for error by the language-based model.

Formally, given extra-linguistic features X_{EL} and language features X_L, the residualized control models are built as follows:

$$\hat{Y} = \alpha \times X_{EL} + \beta \tag{1}$$

$$\epsilon = Y - \hat{Y} \tag{2}$$

$$\epsilon \simeq \gamma \times X_L + \lambda \tag{3}$$

The extra-linguistic control model is parameterized by α weights and the β bias term. ϵ denotes the residual, i.e., the error of the extra-linguistic model. The language-based model aims to predict the ϵ residuals, with γ weights and the λ bias term as parameters.

The motivation for this approach is that extra-linguistic features are more informative and less noisy than the language ones. By exploiting this two-stage learning procedure, the model is biased toward favoring the role of extra-linguistics over language features, which prevents the powerful but rare extra-linguistic features from being lost among thousands of noisy language features.

3.2 Factor Adaptation

Lynn et al. (2017) introduced *user-factor adaptation*, a technique for combining message-level features with user-level information (or *factors*) at the feature level. User-factor adaptation, which is based on the feature augmentation approach for domain adaptation (Daumé III, 2007), uses the extra-linguistic features to transform the language-based features. Each of the language-based features has additional, corresponding features that are a composite of itself and an extra-linguistic factor. In this way, the model is able to capture both factor-specific and factor-general properties of each of the language-based features.

Following the work of Lynn et al. (2017), we use a multiplicative composition function for combining the linguistic and extra-linguistic features.

Instead of using user-level factors, we use extra-linguistic variables obtained at the community level, as described below. More formally, let V_j be a matrix such that:

$$\forall j \in \{0, d\} : V_j = v_j \odot \mathbf{1}^l \tag{4}$$

where d is the number of extra-linguistic factors. With n as the number of data instances, let v_j be a column vector of height n where element $v_{j,i}$ is the score of extra-linguistic feature j for instance i. Having l as the number of language variables, in Eq. 4 for each factor j we make a matrix of size $n \times l$, named V_j, in which every column is equal to v_j, and V_j has the same dimensions as language feature matrix X_L. Now for each factor j we use the Hadamard product to multiply V_j with X_L. In this way each row of X_L will be multiplied by the corresponding row in V_j, which is also equal to the corresponding value in v_j. We therefore can write the factor adaptation as follows:

$$X_A = [V_1 \odot X_L, V_2 \odot X_L, \cdots, V_d \odot X_L] \tag{5}$$

The adapted features together with the original language-based features are used for building a single prediction model:

$$\hat{Y} = \alpha \times [X_L, X_A] + \beta \tag{6}$$

3.3 Residualized Factor Adaptation

Even though both the residualized control and factor adaptation approaches exploit extra-linguistics, they combine these in very different ways. The former does it at the model level by learning different models for different types of features and combining those models together. The latter does it at the data level by first combining both sets of features into a transformed feature set and then learning a single model on the obtained features. In addition, these approaches have different motivations and aim to accomplish different objectives.

These modeling differences suggest that the two approaches could have complementary benefits. Residualized factor adaptation (RFA), our proposed method, inherits the advantages of both the residualized control and adaptation techniques, and is depicted in Fig. 1. There are four main steps:

Step 1: Extra-linguistic control model. We build a regression model solely based on the extra-linguistics, as shown in Equation 1, and then compute the residual error of that model as in Equation 2. This error is ultimately used as the outcome label in the final step of RFA.

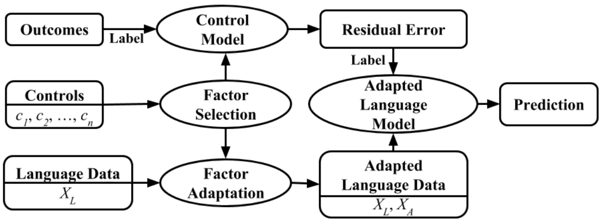

Figure 1: Components of residualized factor adaptation. X_L is language data (topic and n-gram features) and X_A is adapted language data.

Step 2: Factor selection. Adaptation to many factors can increase the model parameters drastically. We explore multiple options for selecting a subset of factors from the available extra-linguistic variables. First, we consider manually selected extra-linguistic factors that are known to influence language use more than others. Second, we use the correlation of the factors with the outcome. Last, we use PCA, an unsupervised method to generate new, lower-dimensional factors from the originals. The purpose of factor selection is to reduce the variance and chance of overfitting.

Step 3: Factor adaptation. We modify the original factor adaptation approach to account for the larger number of factors and features in this task. First, we normalize selected factors by min-max scaling and then multiply the language features by these selected factors as shown in Eq. 5. As we describe later on in Section 4.2, we use n-grams and topics as separate feature sets of language data, so adaptation gives us two corresponding sets of features: adapted n-grams and adapted topics. We then standardize the adapted features using Z-scores and perform feature reduction on each of the four sets of features separately. These reduced features sets are concatenated into a single large set and fed as input to a learning algorithm in the next step.

Step 4: Residual Prediction. The final step, as shown in Eq. 7, is to learn the residual errors of the extra-linguistic control model, using language and adapted language features. Here, we first apply feature selection and reduction on each language feature set: topics, n-grams, adapted-topics and adapted-n-grams. Then we put all of them into a single feature space on which we learn a model to predict the residual error from Step 1.

To produce the final outcome predictions, the predicted error from this model is combined with the predicted outcomes of the extra-linguistic control model from Step 1.

The choice of feature selection is vital for RFA, both due to the fact that it multiplies the number of language features by the number of extra-linguistic features, and because it uses extra-linguistic features at two levels, one separately and one in integration with language features, potentially leading to overfitting. In Section 5.2, we investigate different methods for feature selection to improve RFA's performance.

As Fig. 1 shows, RFA is structured similarly to residualized control. However, residualized control uses language data at its final step, whereas RFA uses both language and adapted language, which is obtained using the factor adaptation technique. This helps RFA to benefit from the advantages of both residualized control and factor adaptation. In other words, RFA combines linguistic and extra-linguistic features on both the feature/data level and the model level. Eq. 7 formulates the RFA method, in which ϵ comes from Eq. 2 and X_A is defined in Eq. 5.

$$\epsilon \simeq \gamma \times [X_L, X_A] + \lambda \tag{7}$$

4 Evaluation Setup

Our task is to predict various community-level outcomes based on publicly available data, including social media and other extra-linguistic data such as socioeconomic and demographic information. We focus on two health-related outcomes: heart disease mortality rate (Eichstaedt et al., 2015) and percent fair/poor health life (Culotta, 2014); one psychology-related outcome: life satisfaction (Schwartz et al., 2013a); and two economy outcomes: Increased real estate price rate and foreclosure rate (Zamani and Schwartz, 2017). Our high-level approach is to train separate regression models for each outcome. For each county, the input is a set of tweets posted by users from that county as well as aggregate values of socioeconomic and demographic variables for the county, including median income, percentage with bachelors degrees and median age. The full list of socioeconomic/demographic variables are given in Section 4.1. The open-source Differential Language Analysis ToolKit was used for the entire analysis pipeline (feature extraction through modeling) (Schwartz et al., 2017)[1].

4.1 Data Set

Our evaluation dataset includes information from three sources: (1) *language data* from Twitter messages, (2) *extra-linguistic data* consisting of 11 socioeconomic and demographic variables, and (3) *outcome data* consisting of 5 county-wise outcomes from 3 categories: Health, Psychology, and Economy.

Our language data can be divided into two groups, (1) for Health and Psychological outcomes and (2) for Economical outcomes. The language data we use for Health- and Psychology-related outcomes was derived from Twitter's 10% random stream collected from July 2009 to February 2015 and includes 1.64 billion tweets (Giorgi et al., 2018)[2]. For Economy outcomes, we used the language data from Zamani and Schwartz (2017). This data was derived from Twitter's 1% random stream collected from 2011 to 2013 and includes 131 million tweets. In both cases, the tweets were mapped to counties based on users' self-reported location strings using the procedure proposed by Schwartz et al. (2013a).

The extra-linguistic data consists of 11 variables used in previous work: 4 *socioeconomic* variables including median income, unemployment rate, percentage of bachelors degrees, and percentage of high school degree, as well as 7 *demographic* variables including median age; percentage: female, black, Hispanic, foreign-born, married; and population density (Census Bureau, 2010). All variables were obtained from the US Census (Census Bureau, 2010), and we henceforth refer to them collectively as *extra-linguistic features*. This dataset is only collected every 10 years, so the 2010 US Census is the most recent dataset for all of the *socioeconomic* and *demographic* variables at the county level.

We consider 5 county-wise measurements as outcomes, 2 health-related (heart disease mortality rate, fair/poor health life), 1 psychological (life satisfaction), and 2 economic (yearly increased real estate price rate, yearly foreclosure rate). Health and psychological data was gathered from the Centers for Disease Control and Prevention (2010b) and contains between 1,630 to 1,749 counties, depending on the outcome. The

[1] Available at `https://github.com/dlatk`

[2] Available at `https://github.com/wwbp`

Domain		Lang.	3 Socio-Demographic Factors					All Factors				
			Controls Only	*Added-Controls*	*RC*	*FA*	*RFA*	*Controls Only*	*Added-Controls*	*RC*	*FA*	*RFA*
Health	HD	0.585	0.423	0.590	0.620	0.628	0.638	0.515	0.597	0.630	0.636	**0.657** *
	FP	0.602	0.434	0.606	0.619	0.647	0.647	0.609	0.632	0.657	0.685	0.680
Psych.	LS	0.214	0.148	0.219	0.292	0.308	0.338	0.326	0.352	0.376	0.353	0.396 *
Econ.	IP	0.245	0.072	0.243	0.266	0.274	0.307	0.240	0.226	0.330	0.344	**0.402** *
	FC	0.153	0.128	0.156	0.197	0.218	0.238	0.160	0.161	0.209	0.240	**0.276** *
	Avg.	0.360	0.241	0.362	0.398	0.415	0.434	0.370	0.394	0.440	0.452	**0.482** *

Table 1: R^2 (variance explained) of residualized factor adaptation (RFA) versus baseline models. Results are shown for 3 hand-picked factors (age, race, education) as well as all factors. RC is residualized control and FA is factor adaptation. Each row is color-coded separately, from red (lowest value) to green (highest values). Bold and * indicate a significant ($p < .05$) reduction in error over the next best model (bold) and FA (*), respectively, according to paired t-tests.

		Lang.	*Controls Only*	*Added-Controls*	*RC*	*FA*	*RFA*
Health	HD	0.765	0.718	0.773	0.794	0.798	**0.811** *
	FP	0.776	0.781	0.795	0.811	0.828	0.825
Psych.	LS	0.463	0.571	0.594	0.614	0.595	0.630 *
Econ.	IP	0.496	0.490	0.476	0.575	0.587	**0.634** *
	FC	0.391	0.401	0.401	0.457	0.490	**0.526** *
	Avg.	0.578	0.592	0.608	0.650	0.659	**0.685** *

Table 2: Pearson-r of residualized factor adaptation (RFA) versus baseline models (for comparison to other work which uses Pearson-r as the accuracy metric). Results are only shown for all factors. RC is residualized control and FA is factor adaptation. Each row is color-coded separately, from red (lowest value) to green (highest values). Bold and * indicate a significant ($p < .05$) reduction in error over the next best model (bold) and over FA (*), respectively, according to paired t-tests.

economic outcomes, which have been used previously in (Zamani and Schwartz, 2017), were gathered for the year 2013 from Zillow[3]. They contain 427 counties' foreclosure rate and 717 counties' increased real estate price rate.

4.2 Baselines

Our baselines consist of a controls-only prediction model and a language-only prediction model.

Controls-only. The controls-only model is a simple regression model trained over all the 11 extra-linguistic features.

Language-only. Building this baseline consists of three main steps: extracting linguistic features, performing feature reduction, and running ridge-regression (Goeman et al., 2016). Our linguistic features are n-gram features (1-3 grams) and topic features which include mentions of 2,000 LDA (Blei et al., 2003) derived topics previously estimated from social media (Schwartz et al., 2013b).

For language data, we first pruned the sparse n-gram features to only include those that were mentioned in at lease a percentage of the counties, then due to the importance of word count in performance of language predictive models(Zamani et al., 2018) we exploit a word count threshold and drop counties with fewer words. Then we run a correlation threshold to only keep the highest correlated features and finally we perform a randomized principal components analysis (RPCA), an approximate PCA based on stochastic re-sampling (Rokhlin et al., 2009). We apply the correlation threshold and RPCA steps for n-grams and topics independently.

For language data associated with health and psychology outcomes, we pruned the sparse n-gram features to only include those that were mentioned in at least 95% of the counties, and used 20,000 as the word count threshold, resulting in 27,250 n-grams total.

With only 1,749 training instances (one per county), feature selection and dimensionality reduction become necessary for avoiding overfitting. We first limit the features to the top 10,000 n-grams with the highest linear relationships to each outcome. As the topic features are more informa-

[3] http://www.zillow.com/research/data/

tive than a single n-gram, we choose to retain all 2,000 topics at this step. Then after performing RPCA we only keep 100 features for each group of ngrams and topcs.

For the language data associated with economy outcomes, we pruned the n-gram features to only include those that were mentioned in at least 10% of the counties, and used 10,000 as the word-count threshold, resulting in 8,897 n-grams across 717 training instances. We use the top 8,000 n-grams and the top 1,500 topic features with the highest linear relationships to each outcome. at the end by applying RPCA we limit the dimension of each feature set to 100.

We compare performances of residualized control, factor adaptation and residualized factor adaptation (RFA). For all these models, we use the same settings as above to generate language features.

5 Results

5.1 Comparison of RC, FA, and RFA

We first compare factor adaptation (FA), residualized control (RC), and residualized factor adaptation (RFA) using three manually selected factors: age, race (percentage of black population), and education (percentage with bachelor's degree) rates. These three factors are often used as "controls" in prior work (Schwartz et al., 2013a; Culotta, 2014; Eichstaedt et al., 2015; Curtis et al., 2018)[4] and also represent examples of demographic and socioeconomic measurements.

In order to ensure a fair comparison, we use the same extra-linguistic features for all models. As mentioned earlier, a naive method is to directly combine the extra-linguistic features with language ones in a single feature set. Here, we also compare this simple model, which we call *added-controls*, with the other three models. In addition, we consider a linear model solely using extra-linguistics, which we call *controls only*. Evaluation is done using 10-fold cross-validation. R^2, or variance explained, is used to measure accuracy.

Table 1 compares results in terms of variance explained, when using the three hand-picked factors vs. using all 11 extra-linguistic factors (Since past work has also used the Pearson-r metric, Table 2 shows the same results for all factors in terms of Pearson-r). As the table shows, FA outperforms controls only, added-controls, and residualized control. RFA does even better and outperforms FA on both the hand-picked factors and when using the entire set of factors. These results demonstrate the complementary nature of the residualized control and factor adaptation approaches and the benefits of combining them.

	No	Separated	Combined	Early
	FS	FS	FS	FS
HD	0.656	**0.657**	0.65	0.639
FP	0.678	**0.68**	0.676	0.661
LS	0.364	0.396	0.391	**0.401**
IP	**0.425**	0.402	0.392	0.336
FC	0.187	**0.276**	0.268	0.241
AVG	0.462	**0.482**	0.475	0.456

Table 3: Comparing R^2 using different methods of feature selection. Outcomes are heart disease (HD), fair/poor health (FP), life satisfaction (LS), increased price (IP), and foreclosure rate (FC). FS stands for feature selection. Bold cells have the highest R^2 for each outcome.

Even though adding controls directly, as in the "added-controls" column, works better than language-only and controls-only models, it is worse than any other model that exploits both language and extra-linguistic data. This motivates the need for combining different types of features in both an additive (residualized control) and multiplicative (factor adaptation) style.

Overall, these results show the power of RFA over the other models. RFA's improvement over FA was statistically significant for 4 out of 5 outcomes, and 3 out of 5 for residualized control. Recall that added-controls, residualized control, FA, and RFA all have access to the same set of information. The gains of RFA over FA show that RFA's structure utilizing residualized control is better suited for combining extra-linguistic and language-only features.

5.2 Feature Selection

Here we investigate the impact of feature selection on the overall performance of RFA. We consider three different combination of adaptation and feature selection, as well as adaptation without any feature selection: (1) *SeparatedFS:* apply feature selection separately on language features and adapted language features; (2) *CombinedFS:* combine language features and adapted language features into one feature set and then apply feature selection; (3) *EarlyFS:* apply feature selection on

[4] Income has also been used frequently but it has been shown to correlate strongly with education rates.

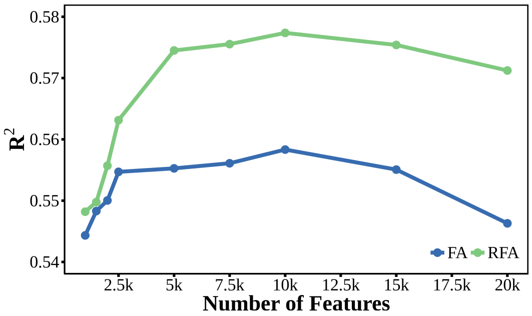

Figure 2: Effect of increasing the number of selected features in univariate feature selection on both factor adaptation (FA) and residualized factor adaptation (RFA) by looking at the average R^2 among health and psychological outcomes. All 11 factors are used in all cases.

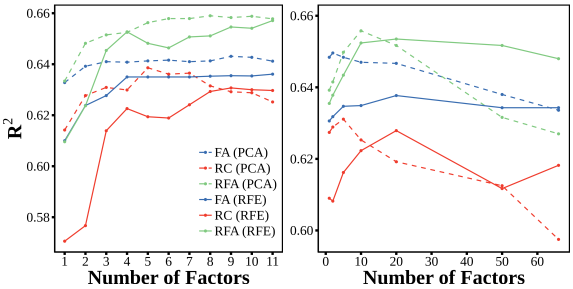

Figure 3: Effect of increasing number of factors on R^2 of residualized factor adaptation (RFA), factor adaptation (FA) and residualized controls (RC) for heart disease outcome. Factors are obtained through Recursive Feature Elimination (RFE) or PCA. Left plot is with original factors, and right plot is with interaction factors (the product from pairing factors).

language features, then apply adaptation on the selected features; and (4) *NoFS:* perform adaptation without any feature selection. Table 3 shows the performance of each method on all 5 outcomes, as well as the the average. SeparatedFS performs better than the others in 3 out of 5 cases, as well as leading the average R^2 across all 5 outcomes. In addition, it produces the most stable results in comparison to the other methods. We therefore use this method for RFA.

We perform another experiment to find the best parameter for the univariate feature selection, that is, the value of k when selecting the k-best n-gram features. Figure 2 shows the results of varying the number of features used for FA and RFA. We report the average R^2 across the 3 health- and psychology-related outcomes. In general, selecting more features leads to better results, though eventually performance does begin to suffer. Recall that our feature selection approach is to first select the k-best n-gram features based on their linear relationship with the outcome, then do a PCA on these k features to obtain a reduced-dimension vector. Even though the feature selection doesn't directly increase the size of our models, it effectively increases the amount of information available to the models, leading to the positive trends we see in Figure 2.

5.3 Increasing Factors and Factor Selection

This experiment has two objectives: first to find out how the number of factors affects performance, and second to find an automated way to select a good subset from the extra-linguistic factors. Here we vary the number of factors from 1 to 11 (i.e. all factors) and compare the effects on RFA, FA, and RC. Factor selection in this experiment is done in two ways, supervised and unsupervised. For the supervised selection we use Recursive Feature Elimination, in which for each k, the least significant factors are recursively dropped until only k factors remain. For unsupervised selection, we use PCA to build k new factors with the highest variance.

The left of Figure 3 shows how the performance of RFA, FA and RC change for heart disease outcome as the number of factors increase, using both PCA and RFE as factor selection methods.

RFA outperforms FA at every factor number, and begins to outperform RC as the number of factors increases. RFA's performance, in general, tends to increase as we add more factors. Using PCA, RFA reaches close to its best performance very quickly, requiring only 5 or 6 factors; adding more factors results in longer runtimes for minimal gain. However, in the case of RFE, using more factors appears to be worthwhile. FA and RC both quickly plateau, or even decline, as more features are added.

Since performance generally improved as more factors were added, we explored adding more factors beyond the 11 that are available to us. To this end we create new factors by multiplying the existing factors with one another. To account for variance in the factor ranges, we first min-max normalize each factor. Then we consider every pair of factors and multiply their normalized values together and re-normalize these new values to create a new factor. This gives a total of 55 new factors in addition to the original 11. We rerun our experiments with this new pool of 66 factors.

The right of Figure 3 shows the results of using this expanded pool of factors. Here, the performance begins to taper off beyond 15 factors for both FA and RFA. Overall, PCA obtains its best performance with only a few factors, but then begins to suffer as more factors are added. RFE, on the other hand, tends to perform worse than PCA initially but remains relatively stable as more factors are added. These newly-created features turned out to be less effective than the original eleven, suggesting that the trade-off in increasing factors via combination is not worthwhile.

Overall, even though reducing the number of factors through PCA-based factor selection could not beat the best accuracy, it is still very competitive. Given the potentially huge number of features obtained through factor adaptation, this slight decrease in performance may be worth potential increases in runtime. RFE-based factor selection, however, helps with neither the runtime nor the performance.

6 Conclusions

Language-based prediction tasks involving communities can benefit from both socio-demographic factors and linguistic features. Because this information comes from different sources and has different distributions, effective mechanisms are needed for combining them. In this paper, we present *residualized factor adaptation*, a method that unites two ways of approaching this problem, one where strong community attributes are augmented (i.e. additive use of factors) with weak but noisy language features, and the other where the contextual differences in language use are mediated via community attributes (i.e. adaptation to community factors). The proposed method effectively combines the complementary benefits of both residualized control and factor adaptation approaches to yield substantial gains over differing community-level prediction tasks across three domains. We see this work as part of a growing need for application-oriented approaches that not only leverage large data effectively by themselves, but do so in the context of other social scientific information that is already available and valuable.

Acknowledgments

This work was supported, in part, by a grant from the Templeton Religion Trust (ID TRT0048). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Centers for disease control and prevention. (2010). underlying cause of death 1999-2010. cdc wonder online database [data set]. retrieved from http://wonder.cdc.gov/ucd-icd10.html.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

United States Census Bureau. 2010. Profile of general population and housing characteristics: 2010 demographic profile data. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&prodType=table.

Deborah A. Cohen, Thomas A. Farley, and Karen Mason. 2003. Why is poverty unhealthy? social and physical mediators. *Social Science & Medicine*, 57(9):1631–1641.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 1–10.

Aron Culotta. 2014. Estimating county health statistics with twitter. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1335–1344. ACM.

Brenda Curtis, Salvatore Giorgi, Anneke E. K. Buffone, Lyle H. Ungar, Robert D. Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H. Andrew Schwartz. 2018. Can twitter be used to predict county excessive alcohol consumption rates? *PLOS ONE*, 13(4):1–16.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 182–188.

Johannes C. Eichstaedt, H. Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Salvatore Giorgi, Daniel Preotiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew

Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jelle Goeman, Rosa Meijer, and Nimisha Chaturvedi. 2016. L1 and l2 penalized regression models.

Oliver L Haimson and Gillian R Hayes. 2017. Changes in social media affect, disclosure, and sociality for a sample of transgender americans in 2016's political climate. In *ICWSM*, pages 72–81.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

Tianran Hu, Ruihua Song, Maya Abtahian, Philip Ding, Xing Xie, and Jiebo Luo. 2017. A world of difference: Divergent word interpretations among people. In *ICWSM*.

Veronica E. 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 1157–1166.

Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, page 182.

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. 2009. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. 2013a. Characterizing geographic variation in well-being using tweets. In *ICWSM*.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes Eichstaedt, and Lyle Ungar. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*.

Christopher Weeg, H Andrew Schwartz, Shawndra Hill, Raina M Merchant, Catalina Arango, and Lyle Ungar. 2015. Using Twitter to measure public discussion of diseases: A case study. *JMIR Public Health and Surveillance*, 1(1).

Mohammadzaman Zamani, Anneke Buffone, and H Andrew Schwartz. 2018. Predicting human trustfulness from facebook language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 174–181.

Mohammadzaman Zamani and H Andrew Schwartz. 2017. Using twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 28–33.