# Domain Adaptation from User-level Facebook Models to County-level Twitter Predictions

**Daniel Rieman**
Positive Psychology Center
University of Pennsylvania
rieman@seas.upenn.edu

**Kokil Jaidka**
Positive Psychology Center
University of Pennsylvania
jaidka@sas.upenn.edu

**H. Andrew Schwartz**
Computer and Information Science
Stony Brook University
has@cs.stonybrook.edu

**Lyle Ungar**
Computer and Information Science
University of Pennsylvania
ungar@cis.upenn.edu

## Abstract

Several studies have demonstrated how language models of user attributes, such as personality, can be built by using the Facebook language of social media users in conjunction with their responses to psychology questionnaires. It is challenging to apply these models to make general predictions about attributes of communities, such as personality distributions across US counties, because it requires 1. the potentially inavailability of the original training data because of privacy and ethical regulations, 2. adapting Facebook language models to Twitter language without retraining the model, and 3. adapting from users to county-level collections of tweets. We propose a two-step algorithm, *Target Side Domain Adaptation* (TSDA) for such domain adaptation when no labeled Twitter/county data is available. TSDA corrects for the different word distributions between Facebook and Twitter and for the varying word distributions across counties by adjusting *target side* word frequencies; no changes to the trained model are made. In the case of predicting the Big Five county-level personality traits, TSDA outperforms a state-of-the-art domain adaptation method, gives county-level predictions that have fewer extreme outliers, higher year-to-year stability, and higher correlation with county-level outcomes.

## 1 Introduction

Social media platforms offer an effective– and widely used– platform for administering surveys to individuals to measure their personality, socioeconomic status, mental and physical well-being, and political orientation, which can then be combined with user posts to build language-based predictive models of user attributes, traits and behaviors. As compared to surveys, language models can be used to assess personality and well-being across communities of the U.S, at a scale not easily achieved by surveys (Eichstaedt et al., 2015). In comparison, Twitter is a more effective tool to mine geographic trends from language, since tweets are publicly accessible, and one in five tweets can be mapped to the county from which they were sent (Schwartz et al., 2013b). However, to our knowledge, there are no tweet-based models of personality which are comparable in accuracy to the Facebook language models of personality.

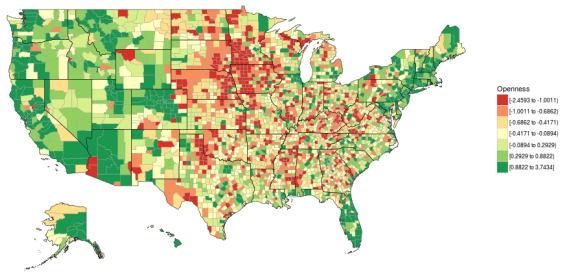


Figure 1: Predictions for county-level openness to experience created by applying a user-level Facebook model for openness on TSDA adjusted Twitter county data

Personality, as measured by the "Big Five" of *openness, conscientiousness, extraversion, agreeableness* and *neuroticism* is known to vary regionally worldwide (Rentfrow et al., 2013) and to cluster geographically in the United States (Rentfrow et al., 2013; Florida, 2002). In this paper, we wish to infer the regional variations of the Big Five Personality traits across the United States, through five language models trained on Facebook posts. We formulate our problem as one of domain adaptation - adapting Facebook models for Twitter's vocabulary, and adapting user-level models for county-level predictions. Figure 1 provides the county-level predictions for the psychological trait of *openness* to experience. At the individual level, *openness* has been found to be correlated with a higher education level and better academic performance (Poropat, 2009). At the regional level, we



*Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 764–773,
Taipei, Taiwan, November 27 – December 1, 2017 

expect to replicate survey results from Rentfrow et al. (Rentfrow et al., 2013), which have demonstrated that regional variations in personality are stable over time, and correlate with key political, economic, social and health metrics.

We combine two forms of target side adaptation in this paper: first, we adapt from Facebook to Twitter; next, we compensate for the variation introduced by the fact that Tweets within each county have significant correlation, leading to spuriously high frequencies of various words in various counties, significantly reducing the predictive accuracy there of the source models. The remainder of the paper first motivates our domain adaptation problem and provides a background on the specific problem of personality prediction. We then situate our method within the field of domain adaptation. Next, we present the TSDA algorithm which, like other popular domain adaptation algorithms, is frustratingly easy (Daumé III, 2007). Finally, we demonstrate that TSDA improves the quality of county-level predictions by (a) removing extreme predictions, (b) improving year-to-year stability, (c) increasing average magnitude of correlations between predicted county-level personality and measured health and well-being metrics with which the personality constructs are known to correlate, and decreasing correlations where correlations are not expected.

### 1.1 The Need for Target Side Adaptation

Applying user-level language models learned on Facebook to make county-level predictions on Twitter poses three main challenges. The first challenge, typically addressed by traditional domain adaptation methods, is the difference in vocabulary between Facebook and Twitter, which motivates the need for domain transfer. For instance, "rt" is one of the most frequent Twitter 'words', but rare on Facebook. Secondly, although typical domain adaptation methods expect the availability of labeled training data, in this case there are new challenges due to the sensitivity of data. In cases where NLP is used to address social science problems, the training data is often unavailable when it comprises personally identifiable information, or when sharing would violate privacy and ethical regulations. Thus, an appropriate unsupervised domain adaptation approach would expect only a trained model – not the original data – to be available for predicting outcomes. A third challenge is that there is a need to disambiguate words which have vastly different frequencies and often entirely different meanings in different counties. The implications of this artifact of the data are obvious when, on reviewing the relative ranking of counties by personality traits, it is observed that the most predictive features for the most- or least-scoring counties, often comprise words which are being used in a different local context (see Table 2). An appropriate domain adaptation approach should account for these local differences, and still generate a generic set of domain adapted features for all counties, rather than 3142 feature sets adapted to each of the counties individually.

We introduce *Target Side Domain Adaptation* (TSDA), an unsupervised method which adapts the target county-level data from Twitter to be more accurately predicted from the source user-level Facebook models *using no labels on the target Tweets or counties and without altering the source-side model*. We call it "Target Side" to emphasize that no retraining of the model is done during the domain adaptation. We assume no labels on the target side, and so only make use of the differences between the source and target distributions of the features. An important assumption in this paper is that differences in the frequencies for the same word among counties, reflects differences in its local meaning. TSDA works particularly well with words, since (a) words vary widely in frequency across domains, and (b) words vary widely in frequency and meaning across domains.

## 2 Background

### 2.1 Personality Traits

We take as our core case study extrapolating personality, as measured in individual level questionnaires, to the 'average' personality for a county. There is a rising research area of "Geographical Psychology" (Rentfrow and Jokela, 2016), which looks at region variations in different psychological traits such as personality, and their correlation with physical and mental well-being.

As a surrogate for large scale surveying, we propose using peoples' social media language to estimate their personality. Such language-based models based on Facebook posts have been built using data from roughly 70,000 people who took personality tests and shared their test results and Facebook posts with researchers (Schwartz et al.,

2013a). These models have proven to be as accurate at estimating personality as estimates from people's friends (Park et al., 2014). However at first blush, such use of Facebook only pushes the problem one level back, as even 70,000 Facebook users give poor coverage of 100 out of over 3,000 US counties. To get good coverage, we shift to a more open social media platform, Twitter.

Twitter is readily available and allows free access to its streaming API. Even though only a few percent of tweets come with latitude and longitude, roughly 20% of the Tweets from the US can still be mapped to their county of origin. Many language-based models of user traits including demographics (Rao et al., 2010; Burger et al., 2011), personality (Schwartz et al., 2013a), socioeconomic status (Preoţiuc-Pietro et al., 2015), popularity (Lampos et al., 2014) and political orientation (Pennacchiotti and Popescu, 2011) have been made from social media language. A number of these models are based on labels of individual tweets (e.g., using Amazon's Mechanical Turk); collecting questionnaire data and the Tweets from the same user is harder, in part due to restrictions on Amazon's terms of use for Mechanical Turk. Facebook requires consenting users to share their data, but while obtaining consent, it is easy to ask users questions to assess their personality, or to ask them to share other data such as their electronic medical records (Smith et al., 2017).

Thus, we face the technical question: How can we take a model trained to predict user attributes such as personality from Facebook language at the individual user level and use it to predict average personality from Twitter language at the county level? This requires a double domain adaptation: firstly from Facebook to Twitter, and secondly from users to counties. This domain adaptation is complicated by the fact that we have virtually no county-level personality measures to use to guide the domain adaptation; it must be unsupervised.

### 2.2 Domain Adaptation Background

Our task can be characterized as domain adaptation, or the closely related transfer learning (Pan and Yang, 2010), where we are adapting from a *source domain*: the words users use on their Facebook posts and associated user labels to a *target domain*: county-level Twitter language, where we want, but do not have, labels on the counties. Most prior work on domain adaptation has focused on the case where some labels are available on both the source and target domains, and is usually done by combining (often in some weighted fashion) training data sets or, less commonly, trained models from the source and target domains (Daume III and Marcu, 2006). Both of these approaches require at least some labeled target data, which we lack. Thus, methods such as EasyAdapt++ (Daumé III et al., 2010), which encourages source and target models to agree on unlabeled data cannot be used here.

In this paper, we have compared our proposed TSDA framework against the Correlation Alignment (CORAL) approach, an unsupervised approach which aims to minimizes domain shift by linearly transforming the covariance matrix of the target distribution to be as similar as possible (under the Frobenius norm) to the source distribution (Sun et al., 2015). It is similar in principle to the study by Daumé and Marcu, which applied Canonical Correlation Analysis (CCA) perform unsupervised machine translation by calculating the cosine similarity of projections in a lower dimensional space (Daumé III and Jagarlamudi, 2011). Transfer Component Analysis (TCA) is a more computationally expensive approach, which exploits the Maximum Mean Discrepancy Embedding (MMDE) metric for comparing the distributions between the source and target domain in the Reproducing Kernel Hilbert Space (RKHS) representation (Pan et al., 2011).

## 3 Target Side Domain Adaptation

Our Target Side Domain Adaptation (TSDA) is a two step process. The first step attempts to minimize the impact of spatially correlated word tokens in the county-level Twitter data by down-scaling the counts of words that are over-represented in some counties. The second step then removes words that have significantly different frequencies between Facebook and Twitter. Note that TSDA does not use any target-side labels. It is instead predicated on the assumption that any large differences in word frequencies between source and target will interfere with the correct generalization; we do not need to know anything about the model in order to do the domain adaptation, instead we use the observed distributions of words on the target side.

Our domain adaptation is motivated by the observation that word frequencies in counties may

have multiple meanings, and that some counties will tend to use an alternative meaning more than others. More formally, we assume that word counts in each county are a mixture of the "true word frequencies" generated by the latent variable being estimated, such as personality combined with county-specific "noise" driven by different word meanings. For example, counties might use the word "jazz" proportionally to how open to experience they are, but a small number of counties (e.g. Salt Lake City) might also use it to refer to a sports team.

Because most words in most counties are generated based on the latent variables of interest (personality), we can compute the distribution of each word across counties (i.e. the distribution on the target side), identify the outliers (words in counties unlikely to come from the main meaning), and then replace them with an imputed value (e.g. the mean frequency for that word).

### 3.1 Step 1: Target-Side Adjustment

As the first step, we adjust the Twitter county-level word frequencies to help mitigate the influence of spatial variations and confounds in word use. This is done by identifying outlier feature values for a given county, and replacing them with the mean feature value across all counties. Extreme values can come from several sources. Common instances of such outliers are: (i) a concert or sports game in a small city, which can lead to disproportionately many mentions of e.g., Bieber or Cowboys. (ii) Some communities have unusually high concentrations of different ethnic groups, who may infuse their nominally English language tweets with Tagalog or Indonesian words. These words, present in small numbers on the source language training set, can significantly skew predictions.

We propose an extremely simple, robust method to address this problem: For each word $w_j$, for the $5\%$ of counties with the largest word frequencies $w_{j,c}$, replace $w_{j,c}$ with the imputed $\bar{w}_j$ and then renormalize each county's word frequencies such that they sum to one.

We attempted using matrix imputation-based values from Singular Value Decomposition (SVD) as in (Troyanskaya et al., 2001) for imputation, as an alternative to using the mean frequency of words; however, we found that it did not make a significant difference on this data set. Also note that in each county, different words are replaced with their imputed values. No single adjustment to the source model is possible; each county effectively gets its own model - because, removing the same 5% of features from all counties would leave in too many harmful features, while removing a feature which is an "outlier" in any county from all other counties would remove too many features that are truly predictive, and also harm model accuracy.

### 3.2 Step 2: Source to Target Adjustment

The second step is to adjust frequencies for words that vary in usage between Facebook and Twitter. As in Step 1, our assumption is that differences in frequency correspond to differences in meaning. Accordingly, for each word in a county, we compute a ratio of its mean frequency for Facebook users, to the mean frequency for Twitter counties. Then, if the ratio for a word lies close to 1.0, word frequencies in the target Twitter counties are replaced with their corresponding means from the source data.

Specifically, we compute

$$\frac{|\bar{w}_j^F - \bar{w}_j^T|}{\bar{w}_j^F + \bar{w}_j^T} > \epsilon$$

Where $\bar{w}_j^F$ is the mean of word frequency $j$ for Facebook users and $\bar{w}_j^T$ is the mean of word $j$ for Twitter counties. In practice we used $\epsilon = 0.8$.

Note that the two domain adaptation steps, although superficially similar, are in fact qualitatively different. In the second step (Facebook to Twitter adaptation) the feature removal is "global", so one could easily remove the features that vary most between the source and target domains and then retrain the source Facebook model without those features. For the first step (cross-county regularization), no such simple retraining is possible.

## 4 Data and Model Description

Facebook user-level models were built using data consisting of 65,896 observations of statuses and personality questionnaire answers. Each user posted at least 1,000 words and answered a set of at least 20 questions to derive a score for each of the Big Five personality traits. Statuses for each of the Facebook users were tokenized, unigram word counts extracted, and converted to term frequencies by dividing the resulting word counts for each

user by that user's total word count. An elastic net regularized linear regression model utilizing a feature selection pipeline described in (Park et al., 2014) was then trained on each of the personality traits.

We used Twitter data comprising the 10% random sample from years 2012-14. We mapped tweets to US counties using the method of Schwartz et al. (2013), which is based on latitude/longitude coordinates, and the self-reported location field when available . Roughly one fifth could be successfully mapped resulting over 150 million geolocated tweets. Only those counties with at least 40,000 words were kept, yielding 2,468 counties for 2012, 2,651 for 2013, and 2,197 for 2014.

## 5 Evaluation of TSDA for Known Outcomes

We first evaluate our predictions by comparing them to five state-level average personality scores (Rentfrow et al., 2013), where we have enough surveys to get a 'ground truth'. The results suggest that TSDA works well when the county-level scores are further aggregated to the state level (Rentfrow and Jokela, 2016). Table 1 shows Pearson $r$ between average state personality predictions (population weighted average of the county predictions) and the 'ground truth'.

The baselines were calculated using predictions with no domain adaptation, with CORAL, and with retrained models on the TCA-transformed features both with and without the feature selection pipeline used in the other models. Recall that CORAL adjusts the source features to the target features using a transformation selected using an $L_2$ (Frobenius norm) loss; such methods work poorly when a small number of extreme values need to be removed.

Additionally, we used this state-level ground truth to test the sensitivity of our method to variation in the step 1 and step 2 parameters. Results showed a statistically significant increase in correlation using 0.8 over 0.9 in step 2 while the variation in the step 1 parameter is less clear. When step 2 uses 0.8, we see openness, conscientiousness and extraversion correlations strengthening with increasing step 1 parameter, agreeableness decreasing, and neuroticism remaining constant. The average increase in performance of TSDA with the range of parameters tested was also statistically significant above the no domain adaptation baseline.

| **Baselines** | | **O** | **C** | **E** | **A** | **N** |
|---|---|---|---|---|---|---|
| No domain adaptation | | .47 | .08 | .49 | .44 | .53 |
| CORAL | | .14 | .19 | .20 | -.16 | .26 |
| TCA no feature selection | | .70 | .26 | -.06 | .64 | .47 |
| TCA feature selection | | .53 | .28 | -.16 | .63 | .48 |
| **TSDA** | | | | | | |
| Step 1 | Step 2 | **O** | **C** | **E** | **A** | **N** |
| 1% | 0.8 | .61 | .09 | .50 | .49 | .57 |
| 2% | 0.8 | .63 | .10 | .50 | .48 | .57 |
| 5% | 0.8 | .69 | .11 | .52 | .45 | .58 |
| 10% | 0.8 | .71 | .13 | .52 | .45 | .57 |
| 1% | 0.9 | .59 | .08 | .49 | .43 | .56 |
| 2% | 0.9 | .61 | .09 | .48 | .42 | .56 |
| 5% | 0.9 | .69 | .10 | .49 | .41 | .57 |
| 10% | 0.9 | .71 | .14 | .48 | .42 | .56 |

Table 1: Pearson $r$ between target side state-level predictions and 'ground truth' state personalities (Rentfrow et al., 2013). The first baseline uses the naive predictions, the second uses the CORAL method, and the third and fourth use the TCA method first without and then with the feature selection pipeline used in the other models. TSDA results are shown for a variety of parameters for steps 1 and 2, but in practice we use 5% and 0.8.

## 6 Unsupervised Validation Method

Validating unsupervised domain adaptation often presents a challenge, when no ground truth is available. We refer to the set of validity criteria developed by social scientists, such as testing *reliability*, the degree to which an assessment tool produces stable and consistent results e.g. over time and *external validity*, the degree to which the results generalize to other settings (ecological validity) and other people (population validity), as well as *predictive validity*, the ability of a stipulated theoretical construct (like Openness to Experience) to predict (correlate with) behavioral or other *criterion*, outcomes that it is theorized to relate to. In this paper, we demonstrate a three-pronged approach to validation, which relies on the fact that the predictions on the many targets (counties) should be (a) normally distributed as they capture natural phenomena, (b) consistent from year to year, (c) correlated with other state- and county-level outcomes such as education, income, health, and happiness which have been measured. Evaluating domain adaptation to a target domain which has no labeled ground truth presents a novel problem. We want, for example to predict the extraversion of different counties, but do not have sufficient data to know, for example, what the mean extraversion scores are of even hundreds of counties.

We propose a three-pronged approach to validating a domain adaptation method on a set of target observations (counties, here) where we have no ground truth:

**1) Are the distributions of the predicted attributes reasonable?** We know, for example, that personality scores are, by construction, Gaussian at the individual level, and that averaging these Gaussians should give a distribution of mean county-level personalities that are Gaussian. However, as shown below, we find our county-level predictions to be far from a normal distribution, with some predictions lying over 10 standard deviations from the mean. We use kurtosis as a concrete metric to assess the impact of domain adaptation on predictions.

**2) Are the estimates stable?** We know that personality at the level of individuals is relatively stable over time; average personality in a county level should be extremely stable from year to year. However, this wasn't the case for our county-level predictions. We use year to year Pearson correlation between predictions as the benchmark metric to measure stability; domain adaptation should increase these correlations.

**3) Do estimates correlate as expected with other outcomes?** We know that personality correlates with many measurable outcomes for which we *do* have county-level measurements such as health and subjective well-being. A good domain adaptation should produce personality predictions which correlate more highly with such outcomes, while reducing unexpected correlations.

## 7 Results

We now demonstrate the utility of TSDA by modeling how personality, as estimated using language on Facebook, can be used to predict county-level average personality from Twitter language.

We use the widely used Five Factor Model (or Big 5) of personality (thousands of papers have been written using it) (McCrae and John, 1992; Digman, 1990), which classifies personality traits into five dimensions: *extraversion* (outgoing, talkative, active), *agreeableness* (trusting, kind, generous), *conscientiousness* (self-controlled, responsible, thorough), *neuroticism* (anxious, depressive, touchy), and *openness* (intellectual, artistic, insightful) all measured using the revised neo personality inventory (Costa and McCrae, 2008).

### 7.1 Qualitative Analysis of TSDA

Recall that the first step of TSDA removes the most different 5% of each feature across counties, replacing them with imputed values, while step 2 removes the words that have the most different frequencies between Facebook and Twitter.

We first look at which words counts are being adjusted. Since step 1 imputes new frequencies for different words for each county, too many words are replaced to show them all. As representative examples, we show in Table 2 the 10 words with the largest change after TSDA step 1 for San Francisco, Salt Lake, and Philadelphia counties. These words, selected *without* looking at the labels, are removed from the model (replacing them with their mean values). We can also look, after the fact, and see which of the removed words had the most influence on the prediction. The 10 words which most affected the predicted openness when they were removed for each of the same three counties are shown in Table 3.

We can also see which words most affected the predicted openness when they were removed from Venango County, PA, one of two counties with extremely low predicted openness: 'yg, ini, sama, ada, lagi, yang, aku, hari, yaa, ga'. These words, lyrics from an Indonesian song, show what can go wrong when models are applied to different domains (counties and years here); a previously rare meme becomes common in one county, making predictions for it highly inaccurate.

| Philadelphia | Salt Lake | San Francisco |
|---|---|---|
| ctfu | utah | - |
| philly | salt | ) |
| philadelphia | lake | de |
| pa | city | que |
| ard | news | la |
| eagles | followers | ™ |
| #philly | #jobs | ’ |
| instagram | ut | el |
| gm | #job | san |
| phillies | solutions | new |

Table 2: Words with the largest change after TSDA step 1 for three cities. We see locations and sports teams, non-English terms, and abbreviations as some of the easily identifiable groups of words adjusted by step 1.

Step 2 imputes the same new word frequencies across all counties and, as shown below in Table 4, gives results that are intuitive. We run step 2 separately here for each of the 3 years of Twitter data, again giving top ranked words that differ slightly from year to year.

| Philadelphia | Salt Lake | San Francisco |
|---|---|---|
| artist (-) | lake (+) | francisco (-) |
| corny (-) | jazz (-) | samsung (+) |
| jersey (+) | slc (-) | woww (+) |
| dickhead (+) | salt (-) | art (-) |
| sheesh (-) | projection (-) | itunes (+) |
| iggy (-) | international (-) | blog (-) |
| shore (+) | canvas (-) | content (-) |
| mic (-) | faucet (-) | vintage (-) |
| imu (+) | masturbation (-) | media (-) |
| eagles (+) | robert (-) | technology (-) |

Table 3: Similar to Table 2, this table looks at features which changed the most, but weights the difference by the Openness model weight. The resulting table shows features whose alteration in TSDA step 1 changed the county-level openness prediction the most. The (+) and (-) indicate if the change had a positive or negative impact on each county's openness prediction respectively.

| Frequent on Facebook | | | Frequent on Twitter | | |
|---|---|---|---|---|---|
| 2012 | 2013 | 2014 | 2012 | 2013 | 2014 |
| paste | =[ | farmville | rt | rt | rt |
| ^ | paste | =] | tweets | 2013 | 2014 |
| 8p | farmville | :[ | #winning | tweets | http |
| % | =] | -- | tf | tf | toned |
| repost | -- | paste | tweeting | tweeting | waistline |
| -- | ^ | ^ | tweet | < | 2013 |
| =[ | finaly | http:// | <<<<< | followers | sheds |
| maths | % | =( | <<<< | tweet | tf |
| ng | 8p | =/ | >>>>> | >>>>> | < |
| eid | mubarak | =p | >>>>>> | >>>>>> | tweets |

Table 4: Features measured most different by TSDA Step 2 on three years of Twitter data. Reported both for features that occurred more on Facebook (the left) and those occurring more on Twitter (the right). All comparison were made with the same source Facebook data.

A consequence of readjusting frequencies of words that differ widely between source and target is that words representing years (e.g. '2015') have very different frequencies in the year that the post or tweet was written. Certain celebrity names behave similarly. Since we often predict on different years than we train, such time-correlated features are frequently dropped out.

### 7.2 Unsupervised Validation

As described above, we validate our model in three ways, measuring prediction 1) normality/kurtosis 2) year to year stability and 3) correlation to other county- and state-level outcomes.

**Measuring Normality** with kurtosis in Figure 2 shows that there were clear problems with the naive method. The personality measures were constructed to be normally distributed, therefore one should expect a sample of personality predictions to be approximately normally distributed with kurtosis near 3. This is clearly not the case with the original predictions, with a three year averaged kurtosis ranging between 8 and 36. The TCA and CORAL baseline predictions also had

| Personality | Original | Step 1 | Step 2 | TSDA |
|---|---|---|---|---|
| O | 35.33 | 3.01 | 27.63 | 2.99 |
| C | 22.22 | 4.10 | 8.47 | 3.60 |
| E | 24.99 | 6.41 | 16.43 | 3.90 |
| A | 8.95 | 2.68 | 10.84 | 2.75 |
| N | 10.56 | 2.61 | 17.97 | 2.71 |

Figure 2: Three year average of county prediction kurtosis

| Personality | Original | Step 1 | Step 2 | TSDA |
|---|---|---|---|---|
| O | .83 | .93 | .82 | .93 |
| C | .76 | .83 | .82 | .85 |
| E | .49 | .57 | .51 | .65 |
| A | .84 | .89 | .82 | .88 |
| N | .81 | .81 | .78 | .81 |

Figure 3: Pearson correlations between 2012 and 2013 predicted county-level personalities.

high kurtosis, with the TCA predictions ranging from 10 to 24 and CORAL from 6 to 69. TSDA however fixes this issue; all the predictions have kurtosis values between 2 and 4 when both steps of TSDA have been applied.

**Year to Year Stability** was assessed by correlating county-level predictions from one year to the following. Figure 3 shows that TSDA increased correlations for predictions that already appeared stable and those with lower correlations. When viewed in conjunction with Figure 2, one can see that the initially high correlations for openness and neuroticism, and to a lesser extent some other predictions, were due to high leverage points.

Spurious (county-specific) words that are stable from year to year can cause both high kurtosis and high year to year correlations. Spurious words that vary year to year, perhaps due to a local meme or news story, cause low year-to-year stability.

**Correlations between Predicted Personality and County Health and Well-being** were calculated. Successful domain adaptation should increase these correlations – or at least drive them towards what is expected from the previous literature. Figure 4 shows the average correlations across three years between observed county-level outcomes and personality predictions both original and post-TSDA word frequencies. We focus first on the overall results. TSDA yields an average increase in correlation magnitude of 11% above the original unadjusted correlations, when averaged over all the personality factors and outcomes listed in the table, as expected. However, some correlations increased due to domain adaptation, while some decreased. We asked a personality psychologist to frame hypotheses for the re-

lationship of personality traits with other county-level outcomes, based on findings from the psychology literature and her considerable experience in the same domain. We tested our findings against these hypotheses, provided in Figure 4 and reflect the expected positive or negative relationship between the five personality traits and income, life satisfaction, mental health, education and income.

| **Hypothesis** | **O** | **C** | **E** | **A** | **N** |
|---|---|---|---|---|---|
| Mentally Unhealthy Days | | - | - | - | + |
| Life Satisfaction | + | + | + | + | - |
| Education | + | + | | | - |
| Income | | + | + | - | - |
| **No domain adaptation** | | | | | |
| Mentally Unhealthy Days | -.03 | -.04 | -.04 | -.06 | .06 |
| Life Satisfaction | .11 | .20 | .01 | .13 | -.17 |
| Education | .32 | .31 | -.03 | .22 | -.30 |
| Income | .18 | .12 | .01 | .13 | -.09 |
| **TSDA** | | | | | |
| Mentally Unhealthy Days | .00 | -.07 | -.04 | -.11 | .10 |
| Life Satisfaction | .12 | .23 | .00 | .13 | -.21 |
| Education | .55 | .35 | -.25 | .07 | -.49 |
| Income | .38 | .11 | -.15 | -.01 | -.25 |

Figure 4: **Top**: Individual-level correlation direction between personality traits and health, well-being, and socioeconomic status measures as predicted by a personality expert, supplemented by meta-analytic findings published in psychology where readily available (Roberts et al., 2007). **Middle**: Pearson *r* between original county-level personality predictions and externally measured metrics. **Bottom**: Pearson *r* between county-level predictions using TSDA and the same externally measured metrics.

We see that our predictions mostly accord with the personality literature, and that domain adaptation often strengthens the correlations that we expected and weakens the correlations that had been predicted from language, but were not expected. In particular, in the original data, there is a predicted strong correlation between agreeableness and education. Given the high kurtosis in agreeableness, this correlation is suspicious. TSDA reduces this correlation from 0.22 to 0.07, much more in line with what would be expected.

Several noteworthy correlations are found in our data, and strengthened by TSDA.

- Openness is known to correlate positively with higher educational attainment. This correlation increases 72% from 0.32 to 0.55 when TSDA is used.

- Conscientiousness is known to correlate positively with life satisfaction, education and income, and negatively with mentally unhealthy days. Our results are consistent with this.

### 7.3 TSDA: Contribution of the two steps

Both steps of TSDA contribute to its performance, as shown in Figures 2 and 3.

*Step 1 (Between County normalization)* appears mostly responsible for the final reduction in kurtosis from TSDA and on its own adds some increased year to year stability. It also significantly increased the county-level predictions' correlations with socioeconomic status, and health and well-being measures and removed the spurious correlation between agreeableness and education which was observed in both the naive application of the model and in results only relying on TSDA step 2.

*Step 2 (Facebook to Twitter normalization)* on its own gave mild improvements in kurtosis (Figure 2), but inconsistent performance in year-to-year stability and external county-level correlations. It is when step 2 is applied after step 1 that the method is truly able to find differing features that, when replaced with the source feature means, provide overall better predictions.

## 8 Conclusion

We have shown that the naive approach to applying Facebook user-level language models to county-level Twitter language has inherent problems due to two separate domain adaptation problems: the differences in Facebook to Twitter word token frequencies, and the spatially correlated terms introduced when aggregating tweets to counties. These problems were discovered when we constructed a list of counties with the most extreme predicted personalities (e.g., 'the 10 most agreeable counties in the US') and found our estimates to be many standard deviations outside what is plausible. We introduce Target Side Domain Adaptation (TSDA), which adjusts the observed word counts in the target (county-level Twitter) domain, leaving the source domain model unchanged, and we propose a set of validation methods based on assessing normality, year-to-year prediction stability, and the correlation of predictions with other outcomes measured on the target counties.

TSDA works particularly well with words, since words vary widely in frequency and meaning across domains and, critically, variations in frequency tend to be associated with differences in meaning. It has the further advantage that it does not require retraining the model; instead, the fea-

ture values passed to the model are readjusted. This could be particularly important when the original training data cannot be shared, for example when it contains personal health data or (as is the case here) private social media data, which are impossible to truly anonymize for sharing.

## Acknowledgments

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048.

## References

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. EMNLP.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment* 2:179–198.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. *ACL 2007* page 256.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 407–412.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, pages 53–59.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research* 26:101–126.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41(1):417–440.

Johannes C Eichstaedt, Hansen 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.

Richard Florida. 2002. The rise of the creative class. *The Washington Monthly* 34(5):15–25.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. EACL.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60(2):175–215.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2014. Automatic Personality Assessment through Social Media Language. *Journal of Personality and Social Psychology* 108(6):934–952.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. ICWSM.

Arthur E Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin* 135(2):322.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An Analysis of the User Occupational Class through Twitter Content. ACL.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. SMUC.

Peter J Rentfrow, Samuel D Gosling, Markus Jokela, David J Stillwell, Michal Kosinski, and Jeff Potter. 2013. Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology* 105(6):996.

Peter J Rentfrow and Markus Jokela. 2016. Geographical psychology: The spatial organization of psychological phenomena. *Current Directions in Psychological Science* 25(6):393–398.

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science* 2(4):313–345.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013a. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE* 8(9).

Hansen 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. 2013b. Characterizing geographic variation in well-being using tweets. In *ICWSM*.

Robert J Smith, Patrick Crutchley, H Andrew Schwartz, Lyle Ungar, Frances Shofer, Kevin A Padrez, and Raina M Merchant. 2017. Variations in facebook posting patterns across validated patient health conditions: A prospective cohort study. *Journal of Medical Internet Research* 19(1):e7.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2015. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547* .

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics* 17(6):520–525.