Human Centered NLP with User-Factor Adaptation

Veronica E. Lynn, Youngseo Son, Vivek Kulkarni
Niranjan Balasubramanian and **H. Andrew Schwartz**
Stony Brook University
Stony Brook, NY
{velynn, yson, vvkulkarni, niranjan, has}@cs.stonybrook.edu

Abstract

We pose the general task of *user-factor adaptation* — adapting supervised learning models to real-valued user factors inferred from a background of their language, reflecting the idea that a piece of text should be understood within the context of the user that wrote it. We introduce a *continuous* adaptation technique, suited for real-valued user factors that are common in social science and bringing us closer to personalized NLP, adapting to each user uniquely. We apply this technique with known user factors including age, gender, and personality traits, as well as latent factors, evaluating over five tasks: POS tagging, PP-attachment, sentiment analysis, sarcasm detection, and stance detection. Adaptation provides statistically significant benefits for 3 of the 5 tasks: up to +1.2 points for PP-attachment, +3.4 points for sarcasm, and +3.0 points for stance.

1 Introduction

Language use is personal. Knowing who wrote a piece of text can help to better understand it. For instance, knowing the age and gender groups of authors has been shown to improve document classification (Hovy, 2015) and sentiment analysis (Volkova et al., 2013).

However, putting people into discrete groups (e.g. age groups, binary gender) often relies on arbitrary boundaries which may not correspond to meaningful changes in language use. A wealth of psychological research suggests people should not be characterized as discrete *types* (or domains) but rather as mixtures of continuous *factors* (McCrae and Costa Jr., 1989; Ruscio and Ruscio, 2000; Widiger and Samuel, 2005).

Here, we ask how one can adapt NLP models to real-valued human *factors* – continuous valued attributes that capture fine-grained differences between users (e.g. real-valued age, continuous gender scores). We refer to this problem as *user-factor adaptation*, and investigate a solution to it in the context of social media, a genre where language is generated by a particularly diverse set of users (Duggan and Smith, 2013). Importantly, *user-factor adaptation* brings us closer to personalized NLP in that with real-valued factors we can now adapt uniquely for each user.

Our approach composes user factor information with the linguistic features, similar to feature augmentation (Daumé III, 2007), a widely used domain adaptation technique which allows for easy integration with most feature-based learning models. Since relevant user information often is not explicitly available, we use a background of tweets from the user to infer user factors. We evaluate our approach over five tasks — POS tagging, PP-attachment, sentiment analysis, sarcasm detection, and stance detection — and with a variety of inferred user factors including (a) *known factors*: age, gender, and personality traits, as well as (b) *latent factors* derived from past user tweets.

Contributions. The main contributions of this work include (a) adaptation based simply on background language (e.g. past tweets; no required *a priori* user knowledge or "domain"), (b) a method for adapting models based on continuous variables, (c) adaptation to other user attributes beyond age and gender (personality and latent factors), and (d) empirical evidence that standard NLP models can often be improved by user-factor adaptation with a range of inferred factors.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1146–1155
Copenhagen, Denmark, September 7–11, 2017.

2 User-Factor Adaptation

User-factor adaptation is especially critical for social media, where content is generated by a diverse user base (Duggan and Smith, 2013). Adaptation requires two components: 1) a user factor representation that captures salient traits indicative of language differences between users, and 2) an adaptation technique that uses this representation to modify learning appropriately.

User factors, even simple ones such as age and gender, may not always be readily available. The messages posted by users, however, are often public and can be used to infer many known linguistically relevant user traits including personality, as well as latent language factors (described next). Given this *background* information about users, the *user-factor adaptation problem* is to learn a single model that is sensitive to both the variations and commonalities in language across different users.

3 User Factors

The first step in our adaptation approach is to create a representation of users that relates to their language use. To this end, we explore two sets of factors: 1) inferred demographics and personality traits, and 2) latent language factors that directly capture language use variations among users.

Different from prior work, we model these human attributes as *real-valued factors*, as is common in psychology literature. Although they may refer to discrete classes such as cluster membership, a factor representation is able to capture more nuanced differences and characteristics that are best understood as a continuum (McCrae and Costa Jr., 1989; Ruscio and Ruscio, 2000; Widiger and Samuel, 2005). This is critical for our goal of moving beyond group-level adaptation toward personalization.

3.1 Demographic and Personality Factors

Many studies have linked language variations with demographic (Argamon et al., 2007; Cheshire, 2005), occupational (Preoţiuc-Pietro et al., 2016) and other psychosocial variables such as personality (Schwartz et al., 2013). We investigate the relevance of a subset of these social variables as user factors for adaptation.

However, we may not have direct access to such information. Unlike the tweets posted by a user, their demographic and personality traits are not always publicly available. We use automatic classification models for obtaining real-valued age and gender estimates (Sap et al., 2014) and personality traits (Park et al., 2015). In addition to being reasonably accurate (e.g. age prediction has a Pearson r of .83 with true age), language based estimation of factor scores may capture linguistic preferences more clearly. For instance, Bamman et al. (2014b) found that *perceived* gender was strongly linked to the gender makeup of a user's social network, and may be a better descriptor of linguistic preferences than self-reported gender.

3.2 Latent Language Factors

We also explore methods to derive latent factors that capture language use similarities and variations across users. The main idea is to derive a latent d-dimensional representation of each user using their background tweets. While there are many choices here, we explore a factorization technique (generative factor analysis), a clustering technique (k-means with TF-IDF), and a hybrid (word2vec with k-means).

Generative Factor Analysis. Factorization methods allow us to build latent representations of users by finding low-rank approximations of the original high-dimensional representations of their text. We use a general method called factor analysis (FA) (Lawley and Maxwell, 1971). Intuitively, FA seeks to capture the variability across correlated variables as a weighted linear combination of a given number of latent dimensions, thus allowing a low-dimensional representation of words[1].

Formally, let $\mathbf{M}_{|\mathcal{U}|\times|\mathcal{V}|}$ denote the user-term matrix, whose entries $\mathbf{M}_{ij}$ indicate the number of times word j is used by user i. FA factorizes this high-dimensional representation into two matrices $\mathbf{F}$ and $\mathbf{L}$ as follows: $\mathbf{M} = \mathbf{FL} + \mathbf{E}$ where $\mathbf{E}$ is an error matrix consisting of residual errors not captured by $\mathbf{FL}$ and where the residual noise is assumed to be Gaussian distributed with zero mean.

Clustering. We also explore commonly used text clustering-based methods to derive *latent factors* from the users' tweets. The idea is to cluster the users based on their tweets. In one case we use TF-IDF based representations, and in the other we use word2vec embeddings (Mikolov et al., 2013).

[1] In this sense FA is a more flexible method than singular value decomposition in that it allows factors to be correlated.

We produce a k-means clustering on this reduced dimensional space to create clusters of users who have similar language use. We derive real-valued factors from these clusters using the distance of the user to the centers of each cluster. Cluster membership yields the discrete representation. Refer to section 5.1 for implementation details.

4 Adaptation Models

Given a factor representation of each user, the adaptation task is to learn a model that is sensitive to both the differences and commonalities across all users. This is similar to the objective for domain adaptation tasks, where the task data is drawn from one or more underlying domains and learning needs to account for both the similarities and differences in the domains. We formulate user-factor adaptation as a domain adaptation technique based on feature augmentation (Daumé III, 2007) but rather than force users into discrete domains, we develop a continuous formulation that allows us to make good use of the real-valued user factors.

Here we first describe a direct discrete formulation of feature augmentation and then describe our proposed continuous formulation.

4.1 Discrete Adaptation

Feature augmentation uses domain information to transform instances into a new augmented space such that instances from the same domain have higher similarity in the augmented space compared to instances from different domains. A learner operating over this augmented space can now learn to model both domain-specific and domain-general influences of the features.

The discrete adaptation method is a direct application of this idea, where the training and test instances are mapped into domains based on some grouping that we induce from the user factors. For example, the user factor *age* induces three discrete *domains*: `low` ($age < 24$), `middle` ($24 < age < 28$), and `high` ($age > 28$).

Given the instance domain mapping, feature augmentation transforms the instances based on their domain. Suppose the original instances have n features and suppose there are d discrete factor classes $(F_1, \cdots, F_d)$ i.e., d domains. Given an instance which is mapped to a factor class F_i, augmentation creates a new feature vector that has $d+1$ feature sets of length n each. The original features are copied over to the first feature set for all instances regardless of their domain. For instances from domain i, the original features are copied over to feature set $i+1$. The other feature sets are zeroes. Table 1 shows some examples of this augmentation strategy for a single instance, $\mathbf{x}$, under different factor class mappings.

User	Factor Classes	Augmented Instance $\Phi(\mathbf{x}, u)$
User 1	F_1	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0 \rangle$
User 2	F_2	$\langle \mathbf{x}, \mathbf{0}, \mathbf{x}, \mathbf{0}, \cdots, 0 \rangle$
User 3	F_1, F_3	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{x}, \cdots, 0 \rangle$
User 4	F_k	$\langle \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0, \mathrm{x} \rangle$

Table 1: Discrete Factor Adaptation: Augmentations of an original instance vector $\mathbf{x}$ under different factor class mappings. With k domains the augmented feature vector is of length $n(k+1)$.

These augmented instances are used for training and testing without any further modifications to the original learning formulation.

4.2 Continuous Adaptation

Discrete adaptation ignores the continuous nature of user factors. Unlike the commonly considered domains, people don't fit neatly into discrete bins. Many psychological studies have shown the ineffectiveness of treating user factors as discrete types (McCrae and Costa Jr., 1989); we expect an adaptation method which does so to be similarly ineffective. For most factors the boundaries for determining classes is unclear, and such arbitrarily-drawn boundaries may not correspond to big changes in language use.

Figure 1 illustrates the advantage of continuous adaptation for a single feature — whether the current instance contains an intensifier — using sarcasm detection as an example. The colored shapes show the feature values for instances from four users, with green squares representing "sarcastic" tweets and yellow circles representing "not sarcastic" ones. The model is unable to distinguish between sarcastic and non-sarcastic tweets in the no adaptation and discrete adaptation case. While discrete adaptation could induce some separability, in this case it fails to account for the variations between differently-aged *over 30* users. On the other hand, if we use features values that are proportional to the actual age, it can result in a better

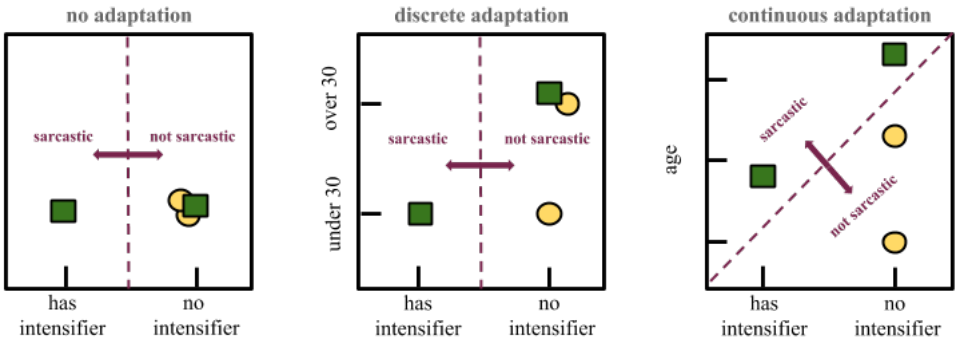


Figure 1: Comparison of feature augmentation under discrete and continuous adaptations. Each shape represents a particular observation (e.g. a tweet) to be classified, each from a different user. The x-axis represents a particular boolean feature: whether the tweet has an intensifier. The y-axis represents how the feature is augmented by the users' age using both discrete adaptation (middle) and continuous adaptation (right). Continuous adaptation allows us to distinguish observations where discrete may not.

separation as shown in the figure.

A compositional function c combines d user factor scores $f_{u,d}$ with original feature values $\mathbf{x}$:

$$\Phi(\mathbf{x}, u) = \langle \mathbf{x}, c(f_{u,1}, \mathbf{x}), c(f_{u,2}, \mathbf{x}), \cdots, c(f_{u,d}, \mathbf{x}) \rangle$$

Thus, a version of each feature exists with and without the factor information integrated. We will explore a simple multiplicative compositional function (i.e., $c(f_{u,d}, \mathbf{x}) = f_{u,d} \cdot \mathbf{x}$) but others can be imagined (e.g. additive, multiplicative with kernel functions).

Multiplicative composition has the property of reducing Φ to discrete adaptation when the factors are binary i.e., $c(f_{u,d}, \mathbf{x}) = \mathbf{x}$ when $u \in F_d$ and $c(f_{u,d}, \mathbf{x}) = 0$ otherwise. As with discrete adaptation, learning then proceeds unmodified with these augmented instances.

The *augmented training data* (train_{aug}) is thus associated with the features x of the tweet, the task labels y, and the user information u. Following the feature augmentation formulation, any supervised learning task of finding a parametrized function h_θ over the original labeled training data can now be specified in terms of the augmented training data along with the transformed instances:

$$\arg\min_{\theta} \sum_{(x,y,u) \in train_{aug}} loss\left(h_\theta\left(\Phi(x,u)\right), y\right)$$

For test instances we apply the same transformation function Φ before prediction.

5 Evaluation

We apply user-factor adaptation to five popular NLP tasks: part-of-speech tagging, prepositional-phrase attachment, sentiment analysis, sarcasm detection, and stance detection. These represent both syntactic and semantic tasks; include some of the key steps in an NLP application pipeline; and use different types of learning formulations including logistic regression, conditional random fields, and support vector machines.

We demonstrate the value of user-factor adaptation on strong baselines for each task. Table 2 provides the specific details for each task including the systems used and their configurations.

5.1 Implementation Details

We learn factors from a user's *background language*, or past tweets[2]. To do so, we collect up to 200 tweets per user; users with fewer than 20 tweets were excluded. Retweets were not included and all tweets were tokenized using the Happier Fun Tokenizer[3].

Demographics and Personality. We derive real-valued demographics and personality scores using the models introduced in section 3.1. For demographics, our model predicts continuous age and a gender score where higher values imply more "femaleness". For personality, these scores represent the Big Five personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg, 1990; McCrae and Costa Jr., 1997). Age, gender, and the five personality dimensions are each a single factor.

[2]Factor inference code is available at: https://stonybrooknlp.github.io/user-factor-adaptation/
[3]https://github.com/dlatk/happierfuntokenizing

Task	POS Tagging	PP-Attachment	Sentiment	Sarcasm	Stance
Output	POS tags	ranked attachments	positive, neutral, negative	sarcastic, not sarcastic	for, against, neutral
System	Owoputi et al. (2013)	variant on Belinkov et al. (2014)	Mohammad et al. (2013)	Bamman and Smith (2015)	Mohammad et al. (2016)
Features	Brown clusters, lexical features	n-grams, Treebank, WordNet	word/char n-grams, lexicon features	all tweet features	word/char n-grams
Learning Alg.	conditional random field	SVM-Rank (Joachims, 2006)	linear-SVM	logistic regression	SVM
Dataset	Owoputi et al. (2013)	Kong et al. (2014) + 986 new tweets	SemEval 2013 (Nakov et al., 2013)	Bamman and Smith (2015)	SemEval 2016 (Mohammad et al., 2016)
Eval	Train/Test, Accuracy	Cross-validation, Accuracy	Train/Test, F1	Cross-validation, F1	Train/Test, F1
Tweets	1544	1319	10339	17084	3021
Users	1541	1319	9917	10966	2349
Instances	22723	2365	10339	17084	3021

Table 2: Overview of the experimental setup for all tasks. Choices were dictated primarily by the literature on top performing systems for each task.

Latent Language Factors. We use three methods to derive latent factors: (1) **tf-idf**: The TF-IDF approach uses unigrams, bigrams, and trigrams occurring in more than 20% but fewer than 80% of documents. (2) **word2vec**: The skip-grams algorithm (Mikolov et al., 2013) was used to produce 50-dimensional word embeddings. (3) **user-embed**: d-dimensional user embeddings from generative factor analysis (Child, 1990) over relative frequencies of n-grams per user-background. The TF-IDF and word2vec representations are then clustered to produce a low-dimensional representation of the users. Each dimension is a single factor. We primarily report results for d=5 for all latent factors, although we explore alternate values in Section 5.3.

Discrete Adaptation. Each user is mapped to a single "domain" per factor. For inferred age, we select three equally-sized domains: $age < 24$, $24 < age < 28$, and $age > 28$. TF-IDF and word2vec define their domains based on cluster membership. Gender, personality, and user embeddings have two domains, above and below the mean, which is done on a per-dimension basis.

Continuous Adaptation. We apply transformations to the raw factor scores before using them for adaptation. For demographic and personality factors, we apply a min-max transformation. Because language often does not vary linearly with age (Pennebaker and Stone, 2003), we additionally use the square root of the predicted age. For the cluster based latent factors, we use the inverse of the Euclidean distance of the user-background from the cluster centroid, amplifying the power of those users who are closest to each cluster. User embeddings from factor analysis are used without any transforms since they naturally produce a Gaussian distribution.

5.2 Results

Table 3 presents the main adaptation results. We compare the performance of adaptation techniques against two baselines: no inclusion of additional factors or adaptation, and models with factors randomly drawn from a Gaussian distribution – a situation requiring learning the same number of parameters as our most augmented models. For the random factor baseline, we take the average performance across five iterations for both discrete and continuous adaptation. To establish significance of difference in error between adaptation results and the no-adaptation baseline, we use permutation testing for stance detection and McNemar's test for the others. Our findings follow. While these conclusions were drawn from our own experiments, we encourage future researchers to see what works best on their own tasks and datasets.

(i) Adaptation improves over unadapted baselines: The results show significant gains with adaptation for PP-attachment (+1.0), sentiment (+1.0), sarcasm (+3.4), and stance (+3.0). Adaptation yields better results for sarcasm and stance, semantic tasks where we'd expect user preferences to be an important factor. While prior studies have shown POS variations across demographic factors (Pennebaker and Stone, 2003; Schwartz

eval measure	pos tagging acc.		pp-attachment acc.		sentiment F1		sarcasm F1		stance F1	
adaptation factors	**disc**	**cont**	**disc**	**cont**	**disc**	**cont**	**disc**	**cont**	**disc**	**cont**
baselines										
no adaptation	91.7	91.7	71.0	71.0	60.6	60.6	73.9	73.9	64.9	64.9
random factors	91.4	91.7	71.0	70.7	59.1	61.1	73.4	74.0	65.5	65.3
user-factor adaptation — known factors										
age	91.5	91.7	69.6	70.8	60.0	**61.4**	**74.9**†	**74.8**†	**66.3**	64.9
gender	91.6	**91.9**	69.7	70.7	**61.0**	**61.0**	**75.0**†	**75.1**†	**66.2**	**65.1**
personality	91.1	91.2	**71.3**	70.2	58.6	**61.2**	**74.3**	**75.6**†	**67.7**†	**66.3**
user-factor adaptation — latent factors										
user embed (d=5)	91.2	90.9	70.7	70.8	59.8	**60.7**	73.9	**77.3**†	64.6	**67.9**†
tf-idf (d=5)	91.4	91.5	70.5	**72.0**†	58.7	**61.6**	73.8	**74.7**†	**66.8**	64.9
word2vec (d=5)	91.6	90.7	70.3	**71.1**	56.30	60.5	**76.4**†	**76.9**†	**67.0**	**66.2**

Table 3: Results of user-factor adaptation across all tasks. Adaptation results are shown in comparison with baseline performance (1) without adaptation and (2) with adaptation using randomly-assigned factors. *disc* denotes discrete adaptation results, and *cont* denotes continuous adaptation results. † indicates statistically significant results at 0.05 level, in comparison to the no-adaptation baseline.

et al., 2013), we hypothesize that the ambiguity in POS reduces greatly when conditioning on local context compared to demographic preferences. This coupled with the ceiling effect in a strong baseline may explain the lack of improvements.

(ii) Continuous is better than discrete: For PP-attachment, sarcasm, and sentiment, continuous adaptation is better than discrete in all but three of the eighteen configurations. Binning people into discrete groups is hard and lossy; using continuous weights helps avoid this issue. Stance, however, is the one task where discrete works better for most factors. As we show in Section 5.4, demographics and personality scores are themselves highly predictive of stances on issues; we believe this makes the simpler binning approach more reliable than the continuous model.

(iii) Both known and latent factors are helpful: Sarcasm benefits from age, gender and personality based adaptations, while stance benefits from personality. The demographic and personality factors do not help PP-attachment. Language factors help all tasks except POS tagging.

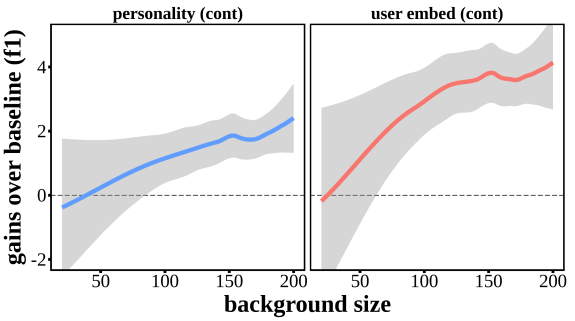
(iv) Latent factors provide best gains: The latent language factors provide the best observed gains for all of the tasks where we saw a significant improvement: PP-attachment, sentiment, sarcasm, and stance. The language factors model users directly in terms of the similarities (and differences) in their entire language use, whereas the inferred demographic and personality factors focus more on a subset of their language as it relates to the particular attribute.

(v) Expanding feature space alone is not enough: One possible explanation for the gains with factors are that the expanded feature space could somehow provide more capacity for learners to generalize. However, adapting to random factors typically lowered results, suggesting that models using more features but not more information naturally take a hit.

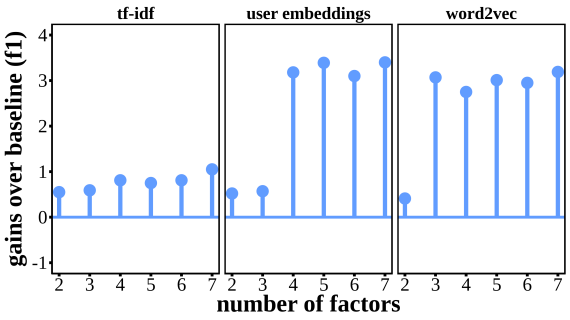
5.3 Background Size and Number of Factors

The amount of background available directly affects the factor measurement, which in turn may impact adaptation effectiveness. Figure 2a shows how varying the background size affects adaptation effectiveness for sarcasm. In general, larger backgrounds lead to bigger gains as expected. Even with small amounts of background (50 tweets) adaptation can provide gains, suggesting that even with imperfect predictions of the user attributes, there is still some benefit to adaptation.

Figure 2b compares how performance varies with the number of latent factors for sarcasm. We see gains for all d sizes we explored. Performance improves with d first and then tapers off; its best is +3.4 at d=5 and 7. As the number of factors increases, there is greater potential for a fine-grained characterization of language use differences. However, this is offset by the increased risk of overwhelming the learner with too many parameters to learn during adaptation. We also find that the impact of number of factors also varies with the type of task (e.g., for PP-attachment we find d=3 gives the best performance of 72.2, a +1.2 gain over the baseline).



(a) Background size effects for cases with large adaptation gains: sarcasm when using personality and user-embedding factors.



(b) Gains over unadapted baseline for sarcasm using TF-IDF, user embeddings, and word2vec with varying number of factors.

Figure 2: Adaptation performance compared against background size and number of factors.

5.4 Factors as Direct Features

One way to use the factors is to add them as direct features to the instances, without adaptation. Table 4 compares how the most beneficial known factor, personality, performs when added directly as a feature to the two tasks where it had the highest impact.

task	base	direct	adapt	best
sarcasm	73.9	**75.6**†	**75.6**†	**77.3**†
stance	64.9	**65.5**	**67.7**†	**67.9**†

Table 4: Effects of including personality scores as direct features, rather than doing adaptation. Other tasks had no benefit from direct features. Best column shows best result achieved with adaptation using any factor.

For sarcasm, adding personality as a direct feature itself leads to a strong improvement on par with using it for adaptation. For stance, however, we see that while there is an improvement over the baseline, it is not as large as that from adaptation. We observed little-to-no improvement for POS tagging, sentiment or PP-attachment when using personality as direct features. This reflects the relative complexity of the relationships between user factors and class labels for each task. Note that while direct features provide benefits, the overall possible gain with adaptation using any factor (shown in best column) is larger.

Including user factors as direct features is beneficial when there is a linear relationship with the class label, such as with gender and sarcasm use. In contrast, user-factor adaptation can capture more complex relationships between user groups and their language expression. Figure 3, for instance, shows a three-way interaction between gender scores, adjective use and sarcasm. Increase in the number of adjectives is a positive indicator of sarcasm for women (high gender scores) but is a negative indicator for men (low gender scores). We observe similar trends for age: phrases such as "thanks" and "love it" tend to be meant sarcastically by younger users and sincerely by older ones. User-factor adaptation can model these interaction relationships when direct features alone may not.

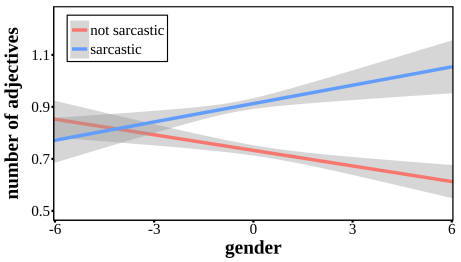


Figure 3: There is a positive correlation between gender and adjective use for sarcastic tweets, and a negative correlation for non-sarcastic tweets. Higher gender scores indicate a greater degree of "femaleness", whereas lower scores represent more "maleness" according to the gender prediction model.

5.5 Comparison of Latent Representations

To understand the advantage of continuous latent adaptation, we look at how well discrete and continuous factor representations capture meaningful information about users. To do so, we select two dimensions from the TF-IDF latent factors for stance detection and examine the extent to which they differentiate users based on their attributes (i.e. age) and posting behavior (i.e. URL use). This is shown in Figure 4. The top row gives the

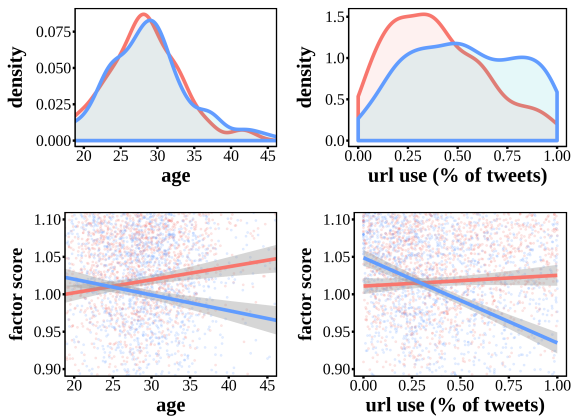


Figure 4: Kernel densities (top) and scatter plots (bottom) of users' age and use of URLs broken down by TF-IDF latent dimension. Colors represent each dimension and are consistent across plots. Lines in scatter plots represent best-fit linear regression. Shaded regions indicate standard error. On the left, discrete latent factors do not seem to distinguish by age (top) but come apart when treated continuously (bottom). On the other hand, discrete and continuous seem to partially capture a dimension of how often someone mentions URLs.

discrete representation: kernel density plots show age and URL use distributions for users binned into the two factor dimensions, shown here in red and blue. The bottom gives the continuous representation: scatter plots show the relationship between age and URL use and the factor score for each dimension.

In the discrete view, age distributions are similar for both factors; there is no apparent relationship between factor membership and age. However, in the continuous view there is a clear negative correlation for age with the factor score for blue and a positive one for red. This indicates that the factors are capturing meaningful information about user age: those with a high factor score for blue tend to be younger, whereas those with a low factor score are older. The reverse is true for red. The URL use shows some difference between the two dimensions in the discrete view, and again we see strong and differing linear relationships with the continuous view.

Overall, the latent factors appear to capture both user attributes and posting behavior, with the continuous version providing additional benefits in characterizing these relationships. The lack of a clear differentiation in the discrete case hints at the difficulty in capturing linguistic variance by splitting users into discrete groups.

6 Related Work

Modeling users has a long history of successful applications in providing personalized information access (Dou et al., 2007; Teevan et al., 2005) and recommendations (Guy et al., 2009; Li et al., 2010; Morales et al., 2012). In contrast, this work models users to better understand their content via language processing tasks following ideas from demographics-aware and domain adaptation.

User-level language variance affects lexical choices (Preoţiuc-Pietro et al., 2016) and even syntactic aspects of language (Johannsen et al., 2015). Such variations can result in demographics-based bias in low-level tasks such as POS tagging (Hovy and Søgaard, 2015) and can also impact high-level applications such as sentiment analysis (Volkova et al., 2013) and machine translation (Mirkin et al., 2015), motivating demographics and personality-based adaptations.

Consequently, recent works have explored demographics-aware NLP (Volkova et al., 2013; Bamman et al., 2014a; Kulkarni et al., 2016; Hovy, 2015; Yang and Eisenstein, 2015). Volkova et al. (2013) propose a gender-aware model and demonstrate superior performance over a gender-agnostic model on the task of sentiment analysis. Bamman et al. (2014a) and Kulkarni et al. (2016) analyze regional linguistic variation using region-specific word embeddings on online social media. Hovy (2015) advances this line of research further and learns group-specific word embeddings, showing improvements over general embeddings on three types of text classification tasks. When author demographics are not available, Yang and Eisenstein (2015) show that learning community-specific embeddings using social networks can help improve sentiment analysis. A similar approach with a social theory-based optimization also showed improvements for sentiment analysis (Hu et al., 2013). For sarcasm detection, historical information about the author and their past context (e.g. entities they discuss) have been shown to be helpful (Bamman and Smith, 2015; Khattri et al., 2015; Rajadesingan et al., 2015).

Our work builds on these ideas and explores the general task of user-factor adaptation. Compared to past work, our method (a) is more general – working with both continuous and discrete factors, (b) uses factors beyond demographics – characteristics like personality are known to influence language beyond demographics (Schwartz

et al., 2013), and (c) only requires a background of language – by using inferred factors from a background of language, we require no *a priori* knowledge of user traits.

7 Conclusion

Language on social media reflects the diversity in its user base and motivates the need for robust models that can handle the resulting variations by user attributes. We have introduced user-factor adaptation, a method to adapt typical supervised language classifiers based on factors of the user authoring the language. Our approach requires nothing more than a background of language by the user and only needs access to the features used by the supervised learner. Since it requires no other modifications to the learner, our approach can be easily applied to many NLP tasks.

To the best of our knowledge, this represents the first work to use the idea of continuous-valued variables for language processing adaptation. Continuous adaptation to a variety of user factors brings us closer to personalized NLP and outperforms discrete adaptation over four different tasks: part-of-speech tagging, preposition-phrase attachment, sentiment analysis, and sarcasm detection. Adaptations with data-driven latent factors produced the largest gains. We see this work as part of a growing trend to put language not just within its document-wide context, but also within the context of the human being that wrote it.

Acknowledgments

This publication was made possible, in part, through the support of a grant from the Templeton Religion Trust – TRT0048. We wish to thank the following colleagues for their annotation help for the PP-attachment task: Chetan Naik, Heeyoung Kwon, Ibrahim Hammoud, Jun Kang, Masoud Rouhizadeh, Mohammadzaman Zamani, and Samuel Louvan.

References

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

David Bamman, Chris Dyer, and Noah A. Smith. 2014a. Distributed representations of geographically situated language. *Proceedings of ACL*, pages 828–834.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Yonatan Belinkov, Tao Lei, Regina Barzilay, and Amir Globerson. 2014. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *TACL*, 2:561–572.

Jenny Cheshire. 2005. Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers. *Journal of Sociolinguistics*, 9(4):479–508.

Dennis Child. 1990. *The Essentials of Factor Analysis*. Cassell Educational.

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

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of WWW*.

Maeve Duggan and Aaron Smith. 2013. Demographics of key social networking platforms. *Pew Research on Social Media*.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216.

Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of RecSys*.

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

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of ACL*.

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of WSDM*.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of KDD*, pages 217–226. ACM.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CONLL*.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Sixth Workshop*

on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), page 25.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of EMNLP*.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Derrick Norman Lawley and Albert Ernest Maxwell. 1971. *Factor analysis as a statistical method*, volume 18. JSTOR.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of WWW*.

Robert R. McCrae and Paul T. Costa Jr. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, 57(1):17–40.

Robert R. McCrae and Paul T. Costa Jr. 1997. Personality trait structure as a human universal. *American Psychologist*, 52(5):509.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of EMNLP*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval-2016*, volume 16.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of SemEval-2013*.

Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In *Proceedings of WSDM*.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934.

James W. Pennebaker and Lori D. Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of AAAI*.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of WSDM*.

John Ruscio and Ayelet Meron Ruscio. 2000. Informing the continuity controversy: A taxometric analysis of depression. *Journal of Abnormal Psychology*, 109(3):473–487.

Maarten Sap, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, David Stillwell, Michal Kosinski, Lyle H. Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of EMNLP*.

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, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS ONE*, 8(9):e73791.

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *SIGIR*.

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

Thomas A. Widiger and Douglas B. Samuel. 2005. Diagnostic categories or dimensions? A question for the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition. *Journal of Abnormal Psychology*, 114(4):494.

Yi Yang and Jacob Eisenstein. 2015. Putting things in context: Community-specific embedding projections for sentiment analysis. *CoRR*, abs/1511.06052.