User Factor Adaptation for User Embedding via Multitask Learning

Xiaolei Huang
Computer Science
University of Memphis
xiaolei.huang@memphis.edu

Michael J. Paul, Robin Burke
Information Science
University of Colorado Boulder
mpaul,robin.burke@colorado.edu

Franck Dernoncourt
Adobe Research
dernonco@adobe.com

Mark Dredze
Computer Science
Johns Hopkins University
mdredze@cs.jhu.edu

Abstract

Language varies across users and their interested fields in social media data: words authored by a user across his/her interests may have different meanings (e.g., cool) or sentiments (e.g., fast). However, most of the existing methods to train user embeddings ignore the variations across user interests, such as product and movie categories (e.g., drama vs. action). In this study, we treat the user interest as domains and empirically examine how the user language can vary across the user factor in three English social media datasets. We then propose a user embedding model to account for the language variability of user interests via a multitask learning framework. The model learns user language and its variations without human supervision. While existing work mainly evaluated the user embedding by extrinsic tasks, we propose an intrinsic evaluation via clustering and evaluate user embeddings by an extrinsic task, text classification. The experiments on the three English-language social media datasets show that our proposed approach can generally outperform baselines via adapting the user factor.

1 Introduction

Language varies across user factors including user interests, demographic attributes, personalities, and latent factors from user history. Research shows that language usage diversifies according to online user groups (Volkova et al., 2013), which women were more likely to use the word *weakness* in a positive way while men were the opposite. In social media, the user interests can include topics of user reviews (e.g., home vs. health services in Yelp) and categories of reviewed items (electronic vs kitchen products in Amazon). The ways that users express themselves depend on current contexts of user interests (Oba et al., 2019) that users may use the same words for opposite meanings and different words for the same meaning. For example, online users can use the word "fast" to criticize battery quality of the electronic domain or praise medicine effectiveness of the medical products; users can also use the words "cool" to describe a property of AC products or express sentiments.

User embedding, which is to learn a fixed-length representation based on multiple user reviews of each user, can infer the user latent information into a unified vector space (Benton, 2018; Pan and Ding, 2019). The inferred latent representations from online content can predict user profile (Volkova et al., 2015; Wang et al., 2018; Farnadi et al., 2018; Lynn et al., 2020) and behaviors (Zhang et al., 2015; Amir et al., 2017; Benton et al., 2017; Ding et al., 2017). User embeddings can personalize classification models, and further improve model performance (Tang et al., 2015; Chen et al., 2016a; Yang and Eisenstein, 2017; Wu et al., 2018; Zeng et al., 2019; Huang et al., 2019). The representations of user language can help models better understand documents as global contexts.

However, existing user embedding methods (Amir et al., 2016; Benton et al., 2016; Xing and Paul, 2017; Pan and Ding, 2019) mainly focus on extracting features from language itself while ignoring user interests. Recent research has demonstrated that adapting the user factors can further improve user geolocation prediction (Miura et al., 2017), demographic attribute prediction (Farnadi et al., 2018), and sentiment analysis (Yang and Eisenstein, 2017). Lynn et al. (2017); Huang and Paul (2019) treated the language variations as a domain adaptation problem and referred to this idea as *user factor adaptation*.

In this study, we treat the user interest as *domains* (e.g., restaurants vs. home services domains) and propose a multitask framework to model language variations and incorporate the user factor into user embeddings. We focus on three online

review datasets from Amazon, IMDb, and Yelp containing diverse behaviors conditioned on user interests, which refer to genres of reviewed items. For example, if any Yelp users have reviews on items of the home services, then their user interests will include the home services.

We start with exploring how the user factor, user interest, can cause language and classification variations in Section 3. We then propose our user embedding model that adapts the user interests using a multitask learning framework in Section 4. Research (Pan and Ding, 2019) generally evaluates the user embedding via downstream tasks, but user annotations sometimes are hard to obtain and those evaluations are extrinsic instead of intrinsic tasks. For example, the MyPersonality (Kosinski et al., 2015) that was used in previous work (Ding et al., 2017; Farnadi et al., 2018; Pan and Ding, 2019) is no longer available, and an extrinsic task is to evaluate if user embeddings can help text classifiers. Research (Schnabel et al., 2015) suggests that the intrinsic evaluation including clustering is better than the extrinsic evaluation for controlling less hyperparameters. We propose an intrinsic evaluation for user embedding, which can provide a new perspective for testing future experiments. We show that our user-factor-adapted user embedding can generally outperform the existing methods on both intrinsic and extrinsic tasks.

2 Data

We collected English reviews of Amazon (health product), IMDb and Yelp from the publicly available sources (He and McAuley, 2016; Yelp, 2018; IMDb, 2020). For the IMDb dataset, we included English movies produced in the US from 1960 to 2019. Each review associates with its author and the *rated item*, which refers to a movie in the IMDb data, a business unit in the Yelp data and a product in the Amazon data. To keep consistency in each dataset, we retain top 4 frequent genres of rated items and the review documents with no less than 10 tokens.[1] We dropped non-English review documents by the language detector (Lui and Baldwin, 2012), lowercased all tokens and tokenized the corpora using NLTK (Bird and Loper, 2004). The review datasets have different score scales. We

normalize the scales and encode each review score into three discrete categories: positive (> 3 for the Yelp and Amazon, > 6 for the IMDb), negative (< 3 for the Yelp and Amazon, < 5 for the IMDb) and neutral. Table 1 shows a summary of the datasets.

2.1 Privacy Considerations

To protect user privacy, we anonymize all user-related information via hashing, and our experiments only use publicly available datasets for research demonstration. Any URLs, hashtags and capitalized English names were removed. Due to the potential sensitivity of user reviews, we only use information necessary for this study. We do not use any user profile in our experiments, except, our evaluations use anonymized author ID of each review entry for training user embeddings. We will not release any private user reviews associated with user identities. Instead, we will open source our source code and provide instructions of how to access the public datasets in enough detail so that our proposed method can be replicated.

3 Exploratory Analysis of User Variations

Language varies across user factors such as user interests (Oba et al., 2019), demographic attributes (Huang and Paul, 2019), social relations (Yang and Eisenstein, 2017; Gong et al., 2020). In this section, our goal is to quantitatively analyze whether the user interests cause user language variations, which can reduce effectiveness and robustness of user embeddings. We approach this by two analysis tasks, first by measuring word feature similarity based on user interests, and second by examining how classifier performance depends on the grouped user interests in which the model is trained and applied.

3.1 Word Usage Variations

Existing methods mainly infer user embeddings from features of text contents (Pan and Ding, 2019). Therefore, word usage variations across user interests will change word distributions and further impact the stability of user embeddings. We aim to test whether there are language variations across the user interests in our datasets and how strong they are.

We consider the word usage as it relates to user embeddings by estimating the overlap of top word

---

[1]The top 4 rated categories of Amazon-Health, IMDb and Yelp are [sports nutrition, sexual wellness, shaving & hair removal, vitamins & dietary supplements], [comedy, thriller, drama, action] and [restaurants, health & medical, home services, beauty & spas] respectively.

Data	Users	Docs	Rated Items	Tokens	Train	Dev	Test
Amazon-Health	11,438	80,592	3,822	127	64,474	8,060	8,061
IMDb	6,089	123,184	642	187	98,548	12,319	12,320
Yelp	76,323	551,695	9,327	152	441,357	55,170	55,171

Table 1: Statistical summary of the Amazon, Yelp and IMDb review datasets. Amazon-Health refers to health-related reviews. Tokens mean the number of average tokens per document. We present the data split for the evaluation task of text classification on the right side.
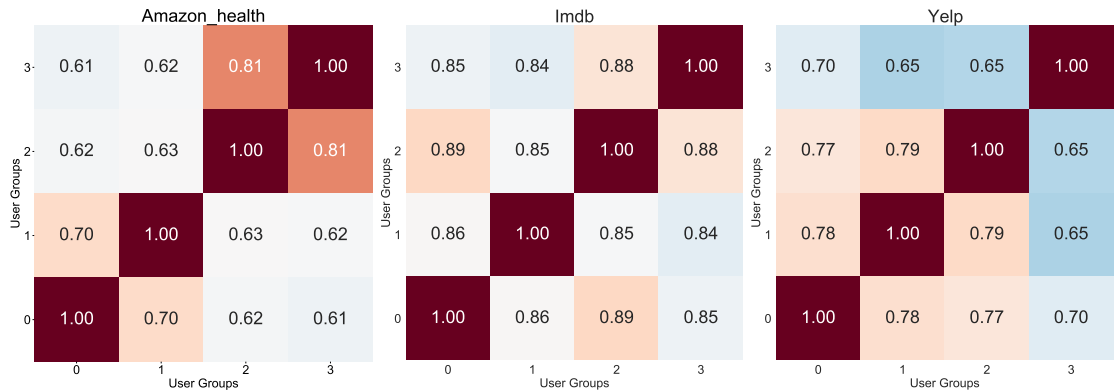

Figure 1: Word feature overlaps between every two user groups. A value of 1 means no variations of top features between two user groups, while values less than 1 indicate more feature variations.

features across the genres of rated items, the categories of reviewed products in Amazon, business units in Yelp and movies in IMDb. To solve data sparsity caused by single user preference, we grouped users and therefore their generated documents according to genres of user reviewed items. We refer to this as genre domains. We build a unified feature vectorizer (Pedregosa et al., 2011) with TF-IDF weighted n-gram features ($n \in \{1, 2, 3\}$), removing features that appeared in less than 2 documents. We rank and select the top 1000 word features for each genre domain by mutual information. We then compute the intersection percentage between every two genre domains: let F_0 is the set of top features for one genre domain and F_1 is the set of top features for the other domain, then the overlap is $|F_0 \cap F_1|/1000$.

We show the results in Figure 1. The overlap varies significantly across genre domains. This indicates that the word usage and its contexts of users change across user interests and preferences. Since the training of user embeddings relies heavily on the language features of users, this suggests that it is important to consider the language variations in user interests for the user embeddings.

3.2 Classification Performance Variations

User embeddings are effective to understand user behaviors in the classification setting (Amir et al.,

2016; Ding et al., 2018). Research has found that combining user and document representations can benefit classification performance (Chen et al., 2016b; Li et al., 2018; Yuan et al., 2019). We explore how the language variations in user interests can affect classification models.

We conduct an analysis by training and testing classifiers that group users by the categories of reviewed items. We first group items and users according to item genres, which can be treated as different domains of user interests. For each domain, we downsampled documents, users and items within each group to match their numbers in the smallest group, so that classification performance differences are not due to data sizes of document, user and item. For each grouped documents, we shuffle and split the data into training (80%) and test (20%) sets. We train logistic regression classifiers with default hyperparameters from scikit-learn (Pedregosa et al., 2011) using TF-IDF weighted uni-, bi- and tri-gram features. We report weighted F1 scores across grouped users and show the results in Figure 2.

We can observe that classification performance varies across the grouped users. Higher performance variations between in- and out- user groups suggest higher user variations and vice versa. If no variations of user language exist, the performance
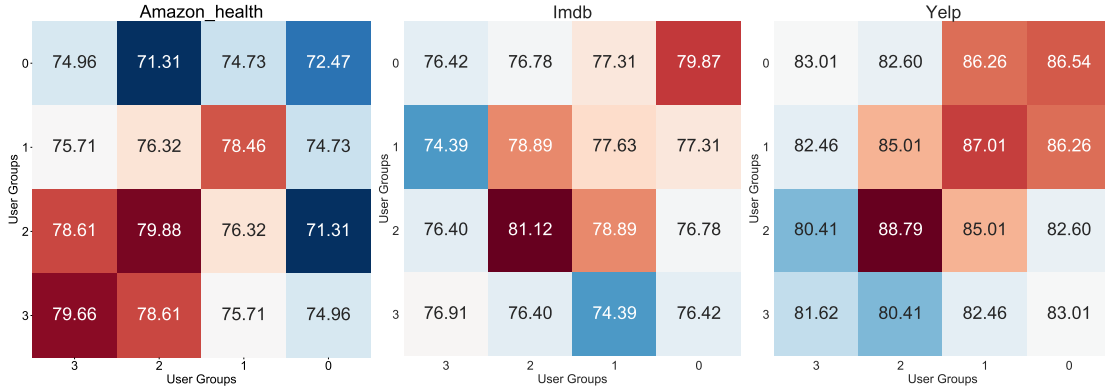
Figure 2: Document classification performance when training and testing on different groups of users. The datasets come from Amazon health, IMDb and Yelp reviews. Darker red indicates better classification performance, while darker blue means worse performance.

of classifiers should be similar across the domains. The performance variations suggest that user behaviors vary across the categories of user interests. We can also observe that classification models generally perform better when tests within the same user groups while worse in the other user groups. This suggests a variability connection between the user interests and language usage, which derives user embeddings.

4 Multitask User Embedding

We present the architecture of our proposed model in Figure 3 on the left. Methods (Pan and Ding, 2019) to train text-based user embedding mainly focus on the user-generated documents while ignoring user factors, the user interests. A close work to ours only trained user embeddings by predicting if users co-occurred with sampled words (Amir et al., 2017). We extend this line of work by adapting user interests into modeling steps. The proposed unsupervised model trains four joint tasks based on the Skip-Gram (Mikolov et al., 2013): word and word, user and word, item and word, and user and item. Note that we do not use the categories of rated items and user interests in our training steps. Then we can optimize the model by minimizing the following loss function:

$$\mathcal{L} = \mathcal{L}(w,w) + \mathcal{L}(u,w) + \mathcal{L}(p,w) + \mathcal{L}(p,u)$$

where w, u, p are the notations of words, users and rated items respectively. Considering the large size of the vocabulary, users and rated items, we approximate our optimization objectives by the negative sampling. Then we can treat each task as a classification problem and calculate loss values by the binary cross-entropy. We present the details of each optimization task as following:

Word and word is a standard way to train Word2vec (Mikolov et al., 2013) models. The prediction task is to predict if the sampled words have co-occurred within the window context. The training process uses the negative sampling to approximate objective function. We choose 5 as the number of negative samples. We keep the top 20,000 frequent words and finally replace the rest with a special token, $<unk>$.

User and word predicts if a user authored the sampled words by the contexts of user posts. The goal is to learn patterns of user language usage from user historical posts. Given a document i, its author u_i and the user's vocabulary $V_{u_i} = \{w_{1,i}, ..., w_{j,i}, ..., w_{n,i}\}$, where n is the number of frequent words authored by the user. Our objective is to minimize the following function:

$\mathcal{L}(u,w) = -\sum_{w_j \in V_{u_i}, w_k \in V, w_k \notin V_{u_i}}[log(\theta(e(u_i) \cdot e(w_j))) + log(\theta(e(u_i) \cdot e(w_k)))]$ where w_j is a negative sample from the whole vocabulary V, $e(u)$ and $e(w)$ are fixed-length user and word vectors respectively, and θ is a sigmoid function to normalize values of dot production. We extend the previous work (Amir et al., 2017) to integrate both local and global user language usage by sampling w_j from a combined token list of both the input document and the user's vocabulary. This can help the model learn contextual information of each user.

Item and word follows the prediction task of user and word to classify if sampled words describe the selected item. This task is to use review documents to train representations of rated items. Then we can have

$\mathcal{L}(p,w) = -\sum_{w_j \in V_{p_i}, w_k \in V, w_k \notin V_{p_i}} log(1 -$
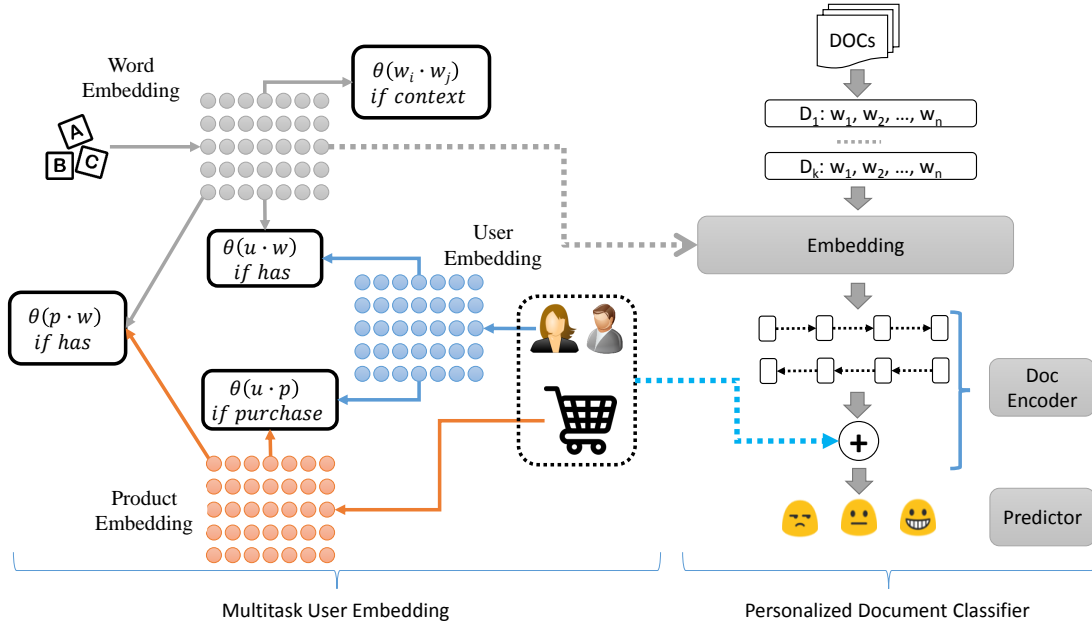
175

Figure 3: Illustrations of User Embedding via multitask learning framework on the left and personalized document classifiers using trained embedding models on the right. The arrows and their colors refer to the input directions and input sources respectively. We use the logos of people, shopping cart and ABC to represent users, reviewed items and word inputs. The $\bigoplus$ is the concatenation operation.

$\theta(e(p_i) \cdot e(w_j))) + log(\theta(e(p_i) \cdot e(w_k)))$ where V_{p_i} is the vocabulary of the rated item p_i and w_k is a negative sample of words. The language can be viewed as a bridge between an interactive relation of user and item, which predicts language usage for both rated items and users.

User and item learns if a user commented on the sampled items. The prediction task aims to adapt latent user factors into the user embeddings. Given a document i, its author u_i and the reviewed item p_i, we can optimize the task by minimizing

$\mathcal{L}(u, p) = -\sum_{p_k \in P, p_k \notin P_{u_i}} log(\theta(e(p_i) \cdot e(u_i))) + log(\theta(e(p_k) \cdot e(u_i)))$ where the P is a collection of all items, the P_{u_i} is a reviewed item and the p_k is a negative sample that the user does not review. The constraints between reviewed items and users can help user embeddings identify language variations across domains of item genres. And in turn, the relation of user and item can help infer item vectors.

For model settings, we used Adam (Kingma and Ba, 2014) for the model optimization with a learning rate of 1e-5. We set the training epochs as 5. The model initializes embedding vectors randomly and learns 300-dimension representations for words, users and reviewed items. We empirically use 5 as the number of negative samples. For the other parameters, we keep the same as the

defaults in the Keras (Chollet and Others, 2015).

5 Experiments

We evaluate the effectiveness of the user factor adapted embedding model by an intrinsic evaluation, user clustering task and an extrinsic evaluation, personalized classification task. The first task aims to measure the purity of clusters with respect to categories of user interests, and the second task uses the document classification as a proxy of qualifying quality of user embeddings. We conduct a qualitative analysis of the user embeddings comparing with our close work (Amir et al., 2017).

5.1 User Clustering Evaluation

The unsupervised evaluation of embedding models focuses on four main categories: relatedness, analogy, categorization and selectional preference (Schnabel et al., 2015). We approach the user embedding evaluation by categorizing users into different clusters. User communities or groups gather users by their interests and behaviors, such as engaging in the same filed of topics (Benton et al., 2016; Yang and Eisenstein, 2017). In our datasets, the user-purchased Amazon products, the user-visited Yelp business units and the user-watched IMDb movies have their item categories. The categories can imply user preferences and inter-

176

ests, and therefore can help evaluate user clusters. In this study, our proposed multitask model learns interactive relations across language, user and item instead of using the item categories. We compare our proposed model with other 5 baseline models:

word2user represents users by aggregating word representations (Benton et al., 2016). We compute a user representation by averaging embeddings of all tokens that were authored by the user. To obtain the word embeddings, for each dataset, we trained a word2vec model for 5 epochs using Gensim (Rehurek and Sojka, 2010) with 300-dimensional vectors.

lda2user generates user representations by applying Latent Dirichlet Allocation (LDA) (Blei et al., 2003) on user documents (Pennacchiotti and Popescu, 2011). We set the number of topics as 300 and leave the rest of the parameters as their defaults in Gensim (Rehurek and Sojka, 2010). We apply the LDA model on each user document to obtain a document vector, and then get a user vector by averaging the vectors of all the user's documents.

doc2user applies paragraph2vec (Le and Mikolov, 2014) to obtain user vectors. We implemented the User-D-DBOW model which achieved the best performance in the previous work (Ding et al., 2017). The implementation keeps parameters with default values in the Gensim (Rehurek and Sojka, 2010). We aggregate each user's documents as a single document. Then the User-D-DBOW model can derive a single user vector from the aggregated document.

bert2user follows a similar process of the lda2user. We use the "bert-base-uncased" pre-trained BERT model for English from the transformers toolkit (Wolf et al., 2019) with default parameter and model settings. After inserting "[CLS]" and "[SEP]" to the beginning and end of each document, the BERT model encodes a document into a fixed-length (768) document vector. We can then generate user embeddings by averaging each user's all document vectors.

user2vec trains user embeddings by predicting word usage by users. We follow the existing work (Amir et al., 2017) but set the user vector dimension as 300.

We use the `SpectralClustering` algorithm from scikit-learn (Pedregosa et al., 2011) toolkit to cluster users into three clustering sizes,

4, 8 and 12. We set the affinity as cosine and leave other parameters as their defaults. To measure cluster quality, we select every two users from the clusters without repetition. We count the user pair as a correct option if two users have overlaps within the same item genre and from the same cluster or if the user pair does not overlap and is from the different clusters. Otherwise, we will count the selection as the wrong option. Therefore, we can have a list of predicted labels and ground truths by using the item genres as a proxy. Finally, we measure the clustering purity by the F1 score.

We present results at Table 2. The results show that our multitask user embedding model outperforms the other baselines by a large portion on the IMDb and Yelp datasets. The improvements suggest the user factor adapted model can understand semantic variations in diverse user interests. The performance of our model and user2vec has similar scores on the Amazon-Health dataset. Comparing to the other two datasets, the Amazon-Health data has more similar topics of review items.

5.2 Personalized Classifier Evaluation

We train three classifiers to evaluate user embeddings on the document classification task. We split each dataset into training (80%), development (10%) and test (10%) sets, as shown in Table 1. The models oversample the minority during the training process. We test the classifiers when they achieve the best performance on the development set. Finally, we report precision, recall and F1 scores using the `classification_report` from scikit-learn (Pedregosa et al., 2011). Figure 3 illustrates personalizing classifiers by concatenating document representations with user embeddings. We compare our proposed model with classifiers using existing user2vec (Amir et al., 2017) and non-personalized classifiers. To ensure a fair comparison, classifiers use the same settings for models with and without user embeddings.

LR. We build a logistic regression classifier using `LogisticRegression` from scikit-learn (Pedregosa et al., 2011). The classifier extracts uni-, bi- and tri-gram features on the corpora with the most frequent 15K features with default parameters.

GRU. We build a bi-directional Gated Recurrent Unit (GRU) (Cho et al., 2014) classifier. We padded documents to the average document length of each corpus. We set the output dimension of GRU as 200

		Amazon-Health			IMDb			Yelp		
		F1@4	F1@8	F1@12	F1@4	F1@8	F1@12	F1@4	F1@8	F1@12
Baselines	word2user	**.929**	.909	.905	.653	.725	.762	.859	.810	.797
	lda2user	.920	**.914**	.900	.696	.726	.761	.849	.839	.832
	doc2user	.873	.891	.901	.660	.725	.748	.836	.828	.826
	bert2user	.871	.896	**.906**	.660	.714	.734	.838	.828	.830
	user2vec	.868	.891	.901	.601	.600	.593	.841	.829	.832
Ours	MTL	.870	.890	.900	**.801**	**.879**	**.884**	**.879**	**.843**	**.838**

Table 2: Performance summary of different user embedding models. We report F1 scores at multiple numbers of clusters. The bold fonts indicate the best performance in each evaluation task.

and apply a dense layer on the output. The dense layer uses ReLU (Hahnloser et al., 2000) as the activation function, applies a dropout (Srivastava et al., 2014) rate of 0.2 and outputs 200 dimensions for final document class prediction. We train the classifier for 20 epochs.

BERT. We implement a BERT-based classifier by HuggingFace's transformers toolkit (Wolf et al., 2019). The classifier loads the "bert-base-uncased" pre-trained BERT model for English, encodes each document into a fixed-length (768) vector and feeds to a linear prediction layer for prediction. We conduct fine-tuning steps for 10 epochs with a batch size of 32 and optimize the model by AdamW with a learning rate of $9e^{-5}$.

We show the performance results in Table 3. Comparing to the baselines, the classifiers personalized by our proposed model generally achieve the best performance across the three datasets. This highlights adapting user factors can help embedding models learn user variations and benefit the classification performance. We can also observe that the personalized classifiers generally outperform the non-personalized classifiers. This indicates personalizing the classifiers with user history boosts classification performance in our study.

5.3 Visualization Analysis

To further evaluate the effectiveness of user embedding models, we map users into a 2-D space using user embeddings and plot them in Figure 4. We group users according to user interests using the domain categories of rated items. To map the 300-d user embeddings, we use the TSNE algorithm from scikit-learn (Pedregosa et al., 2011) to compress the dimension into 2-d vectors. We set the n_component as 2 and leave the other parameters as their defaults in the TSNE. We can observe that the MTL user embedding model shows more cluster-

ing patterns with regard to user interests (categories of reviewed items). This indicates that the unsupervised multitask learning framework can adapt the latent user factors into the user embedding. Users may have multiple interests. In the right plot, we can also find that there is a cluster that mixes with multiple colors on the right bottom.

6 Related Work

User Profiling is a common task in natural language processing. Online generated user texts show demographic variations in the linguistic styles, and the linguistic style variability could be used for predicting user's personality and demographic attributes (Rosenthal and McKeown, 2011; Zhang et al., 2016; Hovy and Fornaciari, 2018; Wood-Doughty et al., 2020; Gjurković et al., 2020; Lynn et al., 2020). The demographic user factors influence how online users express their opinions (Volkova et al., 2013; Hovy, 2015; Wood-Doughty et al., 2017) and show promising improvements in the text classification task (Lynn et al., 2017; Huang and Paul, 2019; Lynn et al., 2019). However, in this work, the goal of modeling user factor is to train robust user embeddings via domain adaptation, rather than the end goal being demographic factor prediction and document classification itself.

Personalized classification generally improves the performance of document classifiers (Flek, 2020). The multitask learning framework has been applied for personalizing document classifiers by optimizing the classifiers on multiple document levels (Benton et al., 2017) or general and individual levels (Wu and Huang, 2016). The social relation can bridge connections between users and generalize classification models across users (Wu and Huang, 2016; Yang and Eisenstein, 2017). For example, (Wu and Huang, 2016) optimizes document

Methods	Amazon-Health			IMDb			Yelp		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LR	.834	.768	.793	.818	.779	.794	.856	.820	.833
LR-u	.841	.777	**.801**	.834	.791	**.807**	.860	.821	.835
LR-up	.838	.771	.796	.833	.791	**.807**	.863	.825	**.838**
GRU	.813	.844	.812	.824	.837	.823	.851	.865	.852
GRU-u	.836	.811	.821	.832	.819	.825	.868	.846	.858
GRU-up	.821	.832	**.825**	.846	.824	**.836**	.876	.864	**.867**
BERT	.866	.822	.840	.852	.809	.826	.866	.825	.840
BERT-u	.863	.812	.831	.858	.818	.833	.872	.843	**.854**
BERT-up	.873	.838	**.851**	.864	.831	**.844**	.880	.839	**.854**

Table 3: Performance scores of document classifiers on the review datasets. '-u' means personalized classifiers using user2vec (Amir et al., 2017) and '-up' indicates personalizing classifiers via our proposed method. We use the bold fonts to highlight the best performance of each classifier on separate datasets.
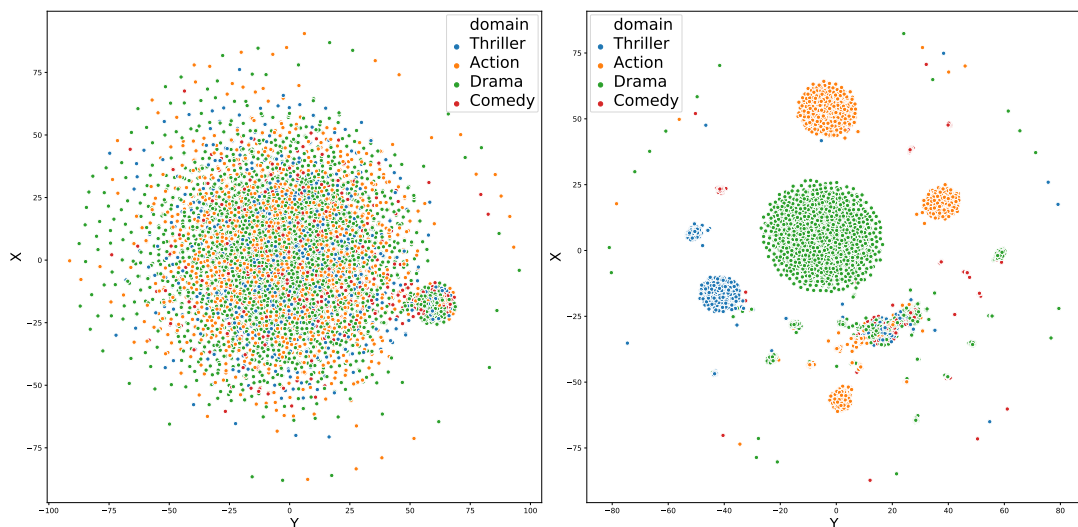

Figure 4: Visualizations of IMDb users colored concerning their interests in 4 movie genres. We plot users using the embeddings from our proposed method (right) and user2vec (Amir et al., 2017) (left). The visualizations of Yelp and Amazon are omitted for reasons of space.

classifiers by two optimization tasks, sentiment classification and user social relation minimization, which allows classifiers to minimize the impacts of user community variations. This work personalizes classifiers in a different way, where we train user embedding models under a multitask learning framework and use the personalized classifiers to evaluate user embedding models.

7 Conclusion

In this study, we have proposed user factor adaptation for building user embedding under a multitask framework. Our analyses show how the user factor causes semantic variations in relation to word usage and document classification, showing that the user factor is rooted in language. We have evaluated the proposed user embedding models in both intrinsic and extrinsic tasks. The user factor adapted model has shown its robustness to language variations in both instrinsic and extrinsic evaluations, learning user representations and personalizing classifiers. We release our source code and instructions of data access at `https://github.com/xiaoleihuang/UserEmbedding`.

Our work in user factor adaptation highlights several future directions to explore. First, our method models latent user factors inferred from user posts. A combination of user embedding and explicit attributes (e.g., demographic factors) may improve model personalization. Second, user behaviors shift over time. A time-adapted user embedding can jointly model temporality and user attributes in online social media and can be extended to other fields, such as public health.

8 Acknowledgement

References

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J. Silva, and Bryon C. Wallace. 2017. Quantifying mental health from social media with neural user embeddings. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321, Boston, Massachusetts. PMLR.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Adrian Benton. 2018. Learning representations of social media users. *arXiv preprint arXiv:1812.00436*.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent {D}irichlet {A}llocation. *Journal of Machine Learning Research*, 3:993–1022.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016a. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659.

Tao Chen, Ruifeng Xu, Yulan He, Yunqing Xia, and Xuan Wang. 2016b. Learning user and product distributed representations using a sequence model for sentiment analysis. *IEEE Computational Intelligence Magazine*, 11(3):34–44.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

François Chollet and Others. 2015. Keras.

T. Ding, W. K. Bickel, and S. Pan. 2018. Predicting delay discounting from social media likes with unsupervised feature learning. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 254–257.

Tao Ding, Warren K. Bickel, and Shimei Pan. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2275–2284, Copenhagen, Denmark. Association for Computational Linguistics.

Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 171–179, New York, NY, USA. Association for Computing Machinery.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the ACL*, pages 7828–7838.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2020. Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*.

Lin Gong, Lu Lin, Weihao Song, and Hongning Wang. 2020. Jnet: Learning user representations via joint network embedding and topic embedding. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 205–213, New York, NY, USA. Association for Computing Machinery.

Richard H R Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of*

the 25th International Conference on World Wide Web (WWW), volume 3, pages 507–517. International World Wide Web Conferences Steering Committee.

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

Dirk Hovy and Tommaso Fornaciari. 2018. Increasing in-class similarity by retrofitting embeddings with demographic information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 671–677, Brussels, Belgium. Association for Computational Linguistics.

Qi Huang, Chuan Zhou, Jia Wu, Mingwen Wang, and Bin Wang. 2019. Deep structure learning for rumor detection on twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Xiaolei Huang and Michael J. Paul. 2019. Neural User Factor Adaptation for Text Classification: Learning to Generalize Across Author Demographics. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*{SEM} 2019)*, volume 4, pages 136–146, Minneapolis, Minnesota. Association for Computational Linguistics, Association for Computational Linguistics.

IMDb. 2020. http://www.imdb.com/interfaces/.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of Machine Learning Research*, volume 32, pages 1188–1196, Bejing, China. PMLR.

Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316, Online. Association for Computational Linguistics.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1260–1272, Vancouver, Canada.

Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, and Masashi Toyoda. 2019. Modeling personal biases in language use by inducing personalized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2102–2108, Minneapolis, Minnesota. Association for Computational Linguistics.

Shimei Pan and Tao Ding. 2019. Social media-based user embedding: A literature review. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6318–6324. International Joint Conferences on Artificial Intelligence Organization.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

181

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *International AAAI Conference on Web and Social Media*.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 763–772.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1014–1023.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring Latent User Properties from Texts Published in Social Media. In *AAAI Conference on Artificial Intelligence (AAAI)*, Austin, TX.

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

Jingjing Wang, Shoushan Li, Mingqi Jiang, Hanqian Wu, and Guodong Zhou. 2018. Cross-media user profiling with joint textual and social user embedding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1410–1420, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zach Wood-Doughty, Michael Smith, David Broniatowski, and Mark Dredze. 2017. How does Twitter user behavior vary across demographic groups? In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 83–89.

Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze. 2020. Using noisy self-reports to predict twitter user demographics. *arXiv preprint arXiv:2005.00635*.

Fangzhao Wu and Yongfeng Huang. 2016. Personalized microblog sentiment classification via multi-task learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3059–3065. AAAI Press.

Zhen Wu, Xin-Yu Dai, Cunyan Yin, Shujian Huang, and Jiajun Chen. 2018. Improving review representations with user attention and product attention for sentiment classification. *arXiv preprint arXiv:1801.07861*.

Linzi Xing and Michael J. Paul. 2017. Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49.

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

Yelp. 2018. Yelp Dataset Challenge.

Zhigang Yuan, Fangzhao Wu, Junxin Liu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2019. Neural review rating prediction with user and product memory. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2341–2344, New York, NY, USA.

Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019. Joint effects of context and user history for predicting online conversation re-entries. In *Proceedings of the ACL*, pages 2809–2818.

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. 2015. Using linguistic features to estimate suicide probability of chinese microblog users. In *Human Centered Computing*, pages 549–559, Cham. Springer International Publishing.

Wanru Zhang, Andrew Caines, Dimitrios Alikaniotis, and Paula Buttery. 2016. Predicting author age from Weibo microblog posts. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 2990–2997.