Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning

Prashanth Vijayaraghavan* MIT Cambridge, MA, USA pralav@mit.edu

Soroush Vosoughi* MIT Cambridge, MA, USA soroush@mit.edu

Deb Roy MIT Cambridge, MA, USA dkroy@media.mit.edu

Abstract

Twitter should be an ideal place to get a fresh read on how different issues are playing with the public, one that's potentially more reflective of democracy in this new media age than traditional polls. Pollsters typically ask people a fixed set of questions, while in social media people use their own voices to speak about whatever is on their minds. However, the demographic distribution of users on Twitter is not representative of the general population. In this paper, we present a demographic classifier for gender, age, political orientation and location on Twitter. We collected and curated a robust Twitter demographic dataset for this task. Our classifier uses a deep multi-modal multi-task learning architecture to reach a state-of-the-art performance, achieving an F1-score of 0.89, 0.82, 0.86, and 0.68 for gender, age, political orientation, and location respectively.

1 Introduction

While the most ambitious polls are based on standardized interviews with a few thousand people, millions are tweeting freely and publicly in their own voices about issues they care about. This data offers a vibrant 24/7 snapshot of people's response to various events and topics.

However, the people using Twitter are not representative of the general US population (Greenwood et al., 2016). Therefore, if one is to use Twitter to understand the public's views on various, it is essential to understand the demographic of the users on Twitter. A robust demographic classification algorithm can also be utilized for detection of non-human account, especially in the context of bots involved in the spread of rumors and false information on Twitter (Vosoughi, 2015).

In this paper, we present a state-of-the-art demographic classifier for Twitter. We focus on four different demographic categories: (a) Gender, (b) Age, (c) Political Orientation and (d) Location. We implement different variants of the deep multi-modal multi-task learning architecture to infer these demographic categories.

2 Features

Our deep multi-modal multi-task learning models (DMT-Demographic Models) use features extracted from the users' Twitter profile (such as name, profile picture, and description), the users following network and the users historical tweets (what they have said in the past). Below, we explain how these features are extracted and used.

2.1 Name

The name specified by users in their profile is mainly used for gender prediction. We used three datasets for gender associations of common names:

- We used the data from US Census Bureau data which contains male and female[1] first names along with frequency of names for the sample male and female population with respect to 1990 census.
- We obtained yearly data for the 100 most popular female and male names between 1960 and 2010 and calculate the overall frequency of a name being used in each list.
- We also have a list of European names and popular first from other countries associated

*The first two authors contributed equally to this work.

[1] https://www2.census.gov/topics/genealogy/1990surnames/

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 478–483
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2076

with gender information[2].

Using the above datasets, we associate each name with a vector of size 2 representing the probability that the name occurs in male list and female list based on frequency available in the datasets.

2.2 Following Network

Network Features can be a signal in prediction of some of the demographic parameters. But it is difficult to utilize the complete list of followers and following information of each and every user due to curse of dimensionality. Therefore, we build an binary vector of size N_{dim} for each user with each index of the vector representing a popular Twitter profiles associated with age, political orientation or location and the value (1, 0) represents if the user is following the profile or not. These profiles are short-listed based on the following techniques:

- We search for user accounts on Twitter for task specific keywords like teenager, 80s, 90s for Age prediction; Democrat, Republican for political orientation and state names for location prediction.
- We take advantage of the data collected from our earlier work (Vijayaraghavan et al., 2016) in processing news stories, classifying named entities into various categories and mapping them to Twitter handles. We use the political personalities mapped onto Twitter to the list of twitter profiles that can potentially act as a signal for our prediction tasks.

Sample Twitter handles associated with each of the tasks are given in Table 1. For gender, the handles were too generic, so we expect that there are inherent latent features that can contribute towards gender prediction based on the shortlisted Twitter handles. We experiment with one or two fully-connected layers and compress the information to a N_{emb}-sized vector.

2.3 Profile Description

The profile description can be really useful to predict all the demographic parameters. Since, GRU is computationally less expensive than LSTM and performs better than standard RNN, we use a gated recurrent network (GRU) (Cho et al., 2014; Chung et al., 2014). At each time step t, GRU unit takes a word embedding x_t and a hidden state h_t as input. The internal transition operations of the GRU are defined as follows:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1} + b^{(z)}) \quad (1)$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1} + b^{(r)}) \quad (2)$$

$$\tilde{h_t} = tanh(Wx_t + r_t \cdot U^h_{t-1} + b^(h) \quad (3)$$

$$h_t = z_t \cdot h_{t-1} + (1 - z_t)\tilde{h_t} \quad (4)$$

where $W^{(z)}, W^{(r)}, W \in \mathbb{R}^{d_h \times d_i}$, $U^{(z)}, U^{(r)}, U \in \mathbb{R}^{d_h \times d_h}$ and $\cdot$ is an element-wise product. The dimensions d_h and d_i are hyperparameters representing the hidden state size and input embedding size respectively. In our experiments, we represent the description as a (a) vector using GRU's final hidden state i.e. the hidden state representation (referred as $D_F \in \mathbb{R}^{d_h}$) at the last time step (b) matrix using all the time steps of hidden state, represented as $D_M \in \mathbb{R}^{L \times d_h}$, where L is the sequence length of the user description.

2.4 Profile Picture

Age and gender prediction can exploit the features extracted from profile picture. We extract dense feature representation from the image using the Inception architecture (Szegedy et al., 2015). Since we deal with multiple tasks, we experiment with two different layers ($pool_3$ and $mixed_10$) representations from the Inception architecture. The output vector sizes are $I_V = 2048$ and $I_M = 64 \times 2048$ respectively.

2.5 Tweets

Finally, we also use tweets for our multitask learning problem. In our experiments, we restrict it to user's recent K tweets. The list of tweets, each of which is word sequence ($S_i = [w^i_1, w^i_2, ..., w^i_N]$), are encoded using a positional encoding scheme as used in (Sukhbaatar et al., 2015). (For a more sophisticated encoding of the tweets, one can use the Tweet2Vec by Vosoughi et al. (Vosoughi et al., 2016), however the algorithm requires a massive training dataset, which might not be available to everyone) For the positional encoding scheme, the sentence representation is computed by

$$P_i = \sum_{j=1}^{N} l_j \cdot w^i_j \quad (5)$$

[2]https://hackage.haskell.org/package/gender-0.1.1.0/src/data/nam_dict.txt.UTF8

Task	Sample Twitter Handles
Age	@80s_Kidz, @The1980sGirl, @60s70sKids, @IL0VEthe80s, @90syears, @The90sLife
Pol-Orien	@realDonaldTrump, @HillaryClinton, @youngdemocrat, @GOP, @NancyPelosi
Location	@california, @UWBadgers, @UtahGov, @UMichFootball, @PureMichigan

Table 1: Sample Twitter handles used to create the network features for each task.

N is the maximum number of words in a sentence and l_j is a column vector with structure

$$l_j^p = (1 - j/N) - (p/q)(1 - 2j/N) \quad (6)$$

where p is the embedding index and q is the dimension of the embedding. The tweet representation obtained from the positional encoding summarizes the word tokens in the sentence. We explore tweet features as (1) a vector by summing up all the K-tweet embeddings $T_q \in \mathbb{R}^q$, (2) a matrix by concatenating all the K-tweet embeddings $T_{Kq} \in \mathbb{R}^{K \times q}$

3 DMT-Demographic Models

Some of the latent information from one task can be useful to predict another task. Therefore, we propose three variants of deep multi-modal multi-task learning demographic models to leverage the multi-modal nature of data. Figure 1 gives an illustration of our proposed models. In this section, we explain the single task output layer followed by the various models.

3.1 Vanilla DMT-Demographic Model

This model takes vector features extracted from various user details (explained in section 2) represented by D_F, T_q, N_{emb}, I_V for description, tweet, network and image features respectively. The feature vectors are concatenated and passed through a fully-connected layer. The output of the fully-connected layer is a compressed latent feature vector of size h. This shared latent vector is given to a task-specific output layer explained in Section 4. For gender prediction task, name features are concatenated with latent vector before feeding it to the output layer.

3.2 Attention-based DMT-Demographic Model

All the modalities do not contribute equally to each of our tasks. Hence, for each task, we concatenate the weighted modal feature representations obtained through attention mechanism and then pass it through a fully-connected layer to get a latent feature vector. Formally, the extracted features vectors represented by D_F, T_q, N_{emb}, I_V are concatenated to get a matrix $M \in \mathbb{R}^{d \times 4}$ where d is the dimension of each feature. If the extracted features are not of the same dimension d, then we introduce a fully-connected layer and transform it to a d-sized vector. The attention over different modal features are computed as follows.

$$\alpha = softmax(W^{(2)} tanh(W^{(1)} M + b^{(1)}) + b^{(2)}) \quad (7)$$

where $\alpha \in \mathbb{R}^{1 \times d}$. We multiply each of the feature vectors by their corresponding α value to get a weighted feature representation. These weighted representation are concatenated before passing it through a fully-connected layer. The latent vector obtained from the fully connected layer is now task-specific and not shared between tasks. The latent vector is given to a task-specific output layer.

3.3 Hierarchical Attention-based DMT - Demographic Model

This model is a slight variant of the previous model. In this model, we introduce another level of attention mechanism over the extracted features. The main intuition behind this approach is to have more attention on individual features based on their importance for a task. For example, certain words like 'male','husband' in user's description might be more suitable for gender prediction than any other task. So we weigh such words higher than the other words in the description during gender prediction task. However, these weights might not be applicable for a location prediction task. Hence, we implement a hierarchical attention mechanism that has task-specific weighted feature extraction followed by task-specific attention over the modalities. The rest of the architecture is similar to the attention-based model.

This model uses the matrix representation associated with each of the features like description (D_M), tweets (T_{Kq}) and profile picture (I_M). However, the network features (N_{emb}) remain unchanged. The attention applied over the extracted features is similar to Equation 7 where the dimen-

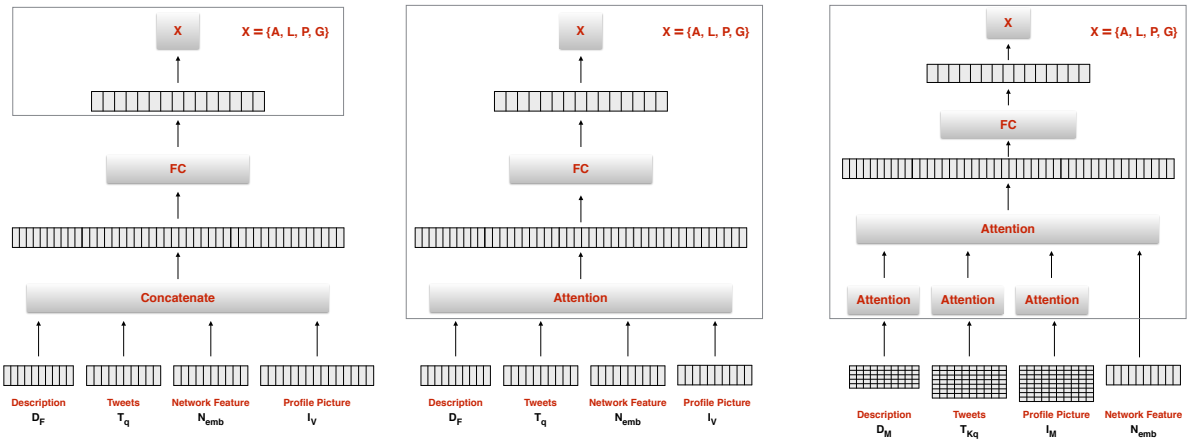

Figure 1: Illustration of variants of the DMT-Demographic Model. Left: Vanilla DMT-Demographic Model; Center: Attention-based DMT-Demographic Model; Right: Hierarchical Attention-based DMT-Demographic Model.

sions of weight parameters are feature-specific. For the sake of convenience, let $\beta^{(F)}$ be the weights similar to α associated with a feature F. For each feature F, we perform a weighted sum over the extracted representation matrix to obtain a vector representation. Let $M^{(F)}$ denote the matrix representation of an extracted feature F, then the vector representation $V^{(F)}$ of the feature F can be computed as follows.

$$V^{(F)} = \sum_{r=1}^{r_F} \beta_r^{(F)} M_r^{(F)} \tag{8}$$

where r_F is the maximum number of rows in the representation matrix $M^{(F)}$ associated with feature F. These vector representations of all the features are fed to layers similar to attention-based DMT model.

It is important to note that all the models incorporate name features with the final latent vector representation for gender prediction task.

4 Output Layer

Given a specific task A, we feed the latent feature vector $h^{(A)}$, obtained after applying any of the explained models, to a softmax layer depending on the classification task. So the task-specific representations are fed to task-specific output layers.

$$\tilde{y}^{(A)} = softmax(W^{(A)} h^{(A)} + b^{(A)}), \tag{9}$$

where $\tilde{y}^{(A)}$ is a distribution over various categories associated with task A.

For each task A, we minimize the cross-entropy of the predicted and true distributions.

$$L^{(A)}(\tilde{y}^{(A)}, y^{(A)}) = \sum_{i=1}^{N} \sum_{j=1}^{C^{(A)}} y_i^{j(A)} log(\tilde{y}_i^{j(A)}) \tag{10}$$

where $y_i^{j(A)}$ and $\tilde{y_i}^j(A)$ are the true label and prediction probabilities for task A, N denotes the total number of training samples and $C^{(A)}$ is the total number of classes associated with the task A. Thus, the parameters of our network are optimized for global objective function given by:

$$\eta = \sum_{A \in X} L^{(A)}(\tilde{y}^{(A)}, y^{(A)}) \tag{11}$$

where X={Age, Location, Political Orientation, Gender}

5 Data Collection & Evaluation

We agglomerated data based on user tweets and their profile description. With access to Twitter API, we were able to get the timeline and profile information of a subset of users. We perform simple analysis of tweets and user description and those that contain phrases like "I'm a girl / woman / guy / man / husband / wife / mother/ father", "I am a democrat / republican / liberal / conservative" or "I support hillary / trump", "Happy 30th birthday to me","I'm 30 years old", "Born in 1980" etc. and their variants are shortlisted. These phrases act as indicators of gender, political orientation and age. For location prediction task, we used a combination of two different Twitter fields to collect

Task	Test Data Size	Majority Classifier (%)
Gender	9,960	53%
Age	6,580	43%
Pol-Orien	5,255	52%
Location	16,956	9%

Table 2: Task-specific details of test data.

data: (a) latitude, longitude from geo-tagged user tweets, (b) Location field in user profile information. The various categories associated with each of the tasks are: (a) Gender: M,F (b) Age: < 30, $30 - 60$, > 60 (c) Political Orientation: Democrat, Republican (d) Location: All states in USA.

In order to avoid selection bias in the dataset collected, we introduce some noise in the training set by randomly removing the terms (from tweet or description) used for shortlisting the user profile. The total size of the training set is 50,859. We evaluate our models on task-specific annotated (mechanical turk) data or data collected based on different phrase indicators from user's tweet or description that was not a part of training set. The details of the test set are given in Table 2. The macro F1-score of different DMT-Demographic models (plus two baseline non-neural network based models) on the test data can be seen in Table 3. Hierarchical-Attention model performs well ahead of the other two models for almost all the tasks. However, the performance of all the models fall flat for location prediction task. Location-specific feature augmentation can be explored to improve its performance further.

6 Related Work

The main distinctions of several of these models with DMT-Demographic models are that (a) most previous literature use only tweet content analysis to predict demographic information (Nguyen et al., 2013) while our model leverages different modals of user information including profile picture, (b) though some of the works use interesting network information they do not leverage other user details as potential signals (Colleoni et al., 2014; Culotta et al., 2015), (c) many of the models involve a lot of feature engineering like extracting location indicative words for geolocation prediction, etc. (Han et al., 2014; Sloan et al., 2015), (d) our model learns shared and task-specific layer parameters as we handle the demographic prediction problem as a multi-task learning problem using different modalities like image (profile picture), text (tweets and user description) and network features (following).

Task	Model	Macro F1
Gender	Random Forrest	0.817
	SVM	0.828
	Vanilla DMT	0.866
	Attention DMT	0.875
	Hierarchical-Attention DMT	**0.890**
Age	Random Forrest	0.724
	SVM	0.733
	Vanilla DMT	0.792
	Attention DMT	0.805
	Hierarchical-Attention DMT	**0.819**
Political Orientation	Random Forrest	0.785
	SVM	0.772
	Vanilla DMT	0.825
	Attention DMT	0.847
	Hierarchical-Attention DMT	**0.859**
Location	Random Forrest	0.668
	SVM	0.665
	Vanilla DMT	0.678
	Attention DMT	0.674
	Hierarchical-Attention DMT	**0.680**

Table 3: Task-specific Macro F1-score for different DMT-Demographic models.

7 Conclusion

In this paper, we presented a state-of-the-art demographic classifier for identifying the gender, age, political orientation and the location of users on Twitter. We also collected and curated a novel Twitter demographic dataset and explored different variants of deep multi-modal multi-task learning architectures, settling on the Hierarchical-Attention DMT as the top performing model, achieving an F1-score of 0.89, 0.82, 0.86, and 0.68 for gender, age, political orientation, and location respectively.

In the future, we intend to use the demographic classifier presented in this paper to study the demographic biases present on Twitter.

References

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* .

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication* 64(2):317–332.

Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the demographics of twitter users from website traffic data. In *AAAI*. pages 72–78.

Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. Demographics of social media users in 2016. http://www.pewinternet.org/2016/11/11/social-media-update-2016/. Accessed: 2017-01-07.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49:451–500.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. " how old do you think i am?" a study of language and age in twitter .

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one* 10(3):e0115545.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. pages 2440–2448.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* .

Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2016. Automatic detection and categorization of election-related tweets. In *Tenth International AAAI Conference on Web and Social Media*.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumors on Twitter*. Ph.D. thesis, Massachusetts Institute of Technology.

Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. 2016. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, pages 1041–1044.