Predicting Twitter User Demographics from Names Alone

Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, Mark Dredze
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218
zach@cs.jhu.edu, noa@cs.jhu.edu, becky@jhu.edu, mdredze@cs.jhu.edu

Abstract

Social media analysis frequently requires tools that can automatically infer demographics to contextualize trends. These tools often require hundreds of user-authored messages for each user, which may be prohibitive to obtain when analyzing millions of users. We explore character-level neural models that learn a representation of a user's name and screen name to predict gender and ethnicity, allowing for demographic inference with minimal data. We release trained models[1] which may enable new demographic analyses that would otherwise require enormous amounts of data collection.

1 Introduction

Social media analysis offers new opportunities for research in numerous domains, including health (Paul and Dredze, 2011), political science (O'Connor et al., 2010), and other social sciences (Gilbert and Karahalios, 2009). Data from social media platforms such as Twitter can yield key insights into population beliefs and behaviors, complementing existing methods such as traditional surveys (Velasco et al., 2014; Dredze et al., 2015). A downside of social media sources is that they often lack traditional demographic information, such as gender, ethnicity, age, and location. Twitter is one of the most popular platforms for research, but its users rarely provide such information.

Numerous existing systems automatically infer missing demographics, such as gender, ethnicity, age and location (Mislove et al., 2011; Burger et al., 2011; Culotta et al., 2015; Pennacchiotti and Popescu, 2011; Rao et al., 2010; Jurgens et al., 2015; Dredze et al., 2013; Rout et al., 2013). Most methods rely on content authored by the user, where words or phrases are strongly associated with specific demographic traits (Al Zamal et al., 2012). Friendship and follower relationships in social networks can also be informative (Chen et al., 2015; Volkova et al., 2014; Bergsma et al., 2013); people tend to be friends with people who live in the same geographic area (Jurgens, 2013) or tend to follow users with similar political orientations (Conover et al., 2011). Culotta et al. (2015) leveraged web traffic data to predict demographics based on who Twitter users follow, e.g. EPSN.com is popular with men, and the @ESPN Twitter account is mostly followed by men.

The principal drawback of these methods is their need for significant data per user, which is often time consuming or expensive to gather. When working with enormous datasets, researchers often avoid demographic analysis altogether, or use limited approaches. For example, a large-scale analysis by Mislove et al. (2011) inferred gender by simply string-matching common names, which failed to label 35.8% of the users studied. Paul and Dredze (2011) tracked flu and allergy symptoms in a dataset of 1.6 million tweets, in which 71% of users had only a single tweet and 97% had 5 or fewer. In a dataset with millions of users, obtaining sufficient content or network data for each user may require prohibitively many Twitter API calls. In production environments, a system may need to make rapid decisions based on a single message, rather than waiting until additional data can be gathered. For these reasons, methods have been proposed for inferring demographics based on the user's name and profile, such as for geolocation, gender, or social roles (Dredze et al., 2013; Osborne et al., 2014; Dredze et al., 2016; Knowles et al., 2016; Volkova et al., 2013; Burger et al., 2011; Beller et al., 2014).

[1] http://bitbucket.org/mdredze/demographer

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 105–111
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

We explore character-level models that learn a low-dimensional representation of a Twitter user's name and screen name, enabling demographic prediction from only a single tweet. Names are a reliable source of demographic information; the name `Sarah` or username `therealjohn` indicate gender, and names like `Carlos` and `Wei` may suggest ethnicity or race. Exact first-name matching has already been proven helpful for demographics inferring, but such methods only work when users use known names (Mislove et al., 2011; Liu and Ruths, 2013; Karimi et al., 2016). Neural models provide the flexibility to learn patterns in character sub-sequences, especially for Twitter names, which are irregular and can contain emojis or special characters. Our model produces more accurate demographic predictions than previous name-based methods, and is competitive with approaches that require more data resources.

2 Models

We hypothesize that character sequences in names are indicative of demographics, and consider models that can learn these correlations from data. Our models encode names and screen names using either convolutional (CNN) and recurrent (RNN) neural networks, which can effectively handle variable-length names. These models convert the tokens of a name into a fixed-length representation, which is then passed through two fully-connected layers to obtain a distribution over the demographic labels.

We searched over a range of model settings:

Single-sequence vs. Multi-sequence Twitter users provide both a name and a screen name; sometimes identical and sometimes completely different. We considered as input either the name only or a concatenation of the name and screen name.

Encoder dimension and depth We considered hidden dimensions ranging from 128 to 1024, both for the number of recurrent cells and for the number of convolutional filters. We additionally considered stacked CNN or RNN components, up to a depth of three layers.

RNN settings Our initial experiments found a Gated Recurrent Unit (GRU) (Cho et al., 2014) cell more effective than Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). We considered both bidirectional and unidirectional RNNs. We evaluated both max-pooling and a learned weighted average[2] to convert the RNN output states into a fixed dimensional embedding.

CNN settings We set the convolution filter width at either two or three. Convolutions had either the ELU activation function (Clevert et al., 2015) or no activation function. All CNN models used max-pooling to reduce the convolution outputs to a fixed dimension. The stacked CNN models used an exponentially increasing dilation rate at each layer (Yu and Koltun, 2015).

Training details We trained all models using cross entropy loss and the Adam optimizer (Kingma and Ba, 2014), using a learning rate of 0.001 and gradients clipped at 10 (Pascanu et al., 2013). Each character in the vocabulary was embedded into a 128-dimension space. Models were implemented in Tensorflow (Abadi et al., 2016).

3 Data and evaluation

We collect data from past work to conduct experiments on gender and ethnicity prediction tasks. We consider both Twitter data and auxiliary government-provided data to train our models, but always evaluate using just Twitter data for the development and test sets. We split the Twitter data into training (60%), development (20%) and test sets (20%). For each task, we use the datasets described below to construct three types of training datasets: only auxiliary data, only Twitter data, and auxiliary plus Twitter data. When using both auxiliary and Twitter data, we train on the auxiliary data until the Twitter development accuracy begins to decline, then switch to training on the Twitter data.

Gender: We consider gender classification as a binary[3] prediction task between men and women, following past work.

Twitter: The dataset created by Burger et al. (2011) (and processed and released by Volkova et al. (2013)[4]) provides us with 58,046 users, 30,364 female and 27,682 male. These labels were

[2] We use the Tensorflow seq2seq implementation of Bahdanau et al. (2014) attention, to convert the sequence of RNN states to a single time-step 'sequence' for classification.

[3] A fuller consideration of of gender identity on Twitter is needed, but is outside the scope of this work.

[4] `http://cs.jhu.edu/~svitlana/data/data_emnlp2013.tar.gz`

obtained for Twitter accounts which linked to a blog in which the author included gender.

Auxiliary Data: We use gender-labeled name data from the Social Security Administration[5] which contains 68,457 unique first names and their co-occurrence with gender. We assigned each name its majority gender label.

Ethnicity: There is limited available training data for race and ethnicity. Due to the large class imbalances in available data, we consider two separate ethnicity tasks. First, we predict Caucasian vs. African-American (2-way), which offered the most data per class and a larger body of past work (Volkova and Bachrach, 2015; Pennacchiotti and Popescu, 2011). Second, we predict Caucasian vs. African-American vs. Hispanic/Latino (3-way) as a more difficult task following Culotta et al. (2015).

Twitter: From the dataset created by Culotta et al. (2015) we collect 407 of the original 770 users[6]: 215 Caucasian, 117 Hispanic/Latino, and 75 African-American. The labels were obtained by manual annotation.[7] Culotta et al. estimated inter-annotator agreement at 80%. From Volkova and Bachrach (2015) we collect 3,862 users of the original 5,000: 1,912 Caucasian, 360 Hispanic/Latino, and 1,309 African American (and 281 other). The labels were obtained by crowdsourced annotations of users' profiles, with a reported Cohen's κ of 0.71.

Auxiliary Data: We use ethnicity-labeled name data from the North Carolina Board of Elections, [8] which contains millions of names labeled with race (White, Black, and five other labels) and ethnicity (Hispanic/Latino, not, or undesignated). We combine race and ethnicity labels into our three classes (Caucasian, African-American, Hispanic/Latino).

3.1 Baselines

For each task we compare our best neural models against two baselines representing prior work: a name-only method and a user content method.

SVM: Knowles et al. (2016) predicts gender with a linear SVM trained on character n-gram features extracted from Twitter users' names. We used the authors' released implementation.

Content: Volkova and Bachrach (2015) predicts gender and ethnicity with a logistic regression classifier trained on the unigrams in the 200 most recent tweets of each user. We used our own implementation, but were unable to test on the exact same data in the original paper. When we evaluated our implementation, our AUC scores were 6-12% lower than those reported by the authors. This difference may be due to changes in the datasets as we have different tweets, fewer users, and different splits.

4 Results

Table 1 shows results on the test data for the best-performing CNN and RNN architecture on each task, with and without auxiliary data. Table 2 shows the results on dev data for each architecture, including results split by name inputs and name plus screen name inputs. We used the dev set performance to pick which architectures to evaluate on the test data, and early-stopping on dev data to get final test scores.

The (macro) F1 score is calculated as the harmonic mean of the average class precision and recall, across each class. [9] While F1 is usually quite similar to accuracy in 2-class comparisons, they diverge in the ethnicity 3-way comparison.

Our models had significantly[10] higher accuracy than the SVM baseline on both gender and ethnicity tasks. While the content baseline outperforms our best models on all tasks, it requires far more data per user.

The use of auxiliary data produced ambiguous results: it greatly helped the SVM model on the gender task, but appeared to hurt performance for all models on the ethnicity tasks. A possible explanation is that, because the SVM only considered simple n-gram features, the informative n-grams for gender are relatively consistent across Census and Twitter names. The neural models, however, learn much more complicated features, and the relevant features

[5] https://www.ssa.gov/OACT/babynames/names.zip

[6] Many users are no longer available on Twitter.

[7] While most accounts correspond to a single individual, some accounts represent entities for which "ethnicity" is not well-defined, but were labeled regardless.

[8] http://dl.ncsbe.gov/index.html?prefix=data/

[9] Knowles et al. (2016) defines F1 as the harmonic mean of accuracy and *coverage*, and thus our F1 scores are substantially lower.

[10] Using a two-proportion z-test, our models outperform the SVM on gender, with $p < 0.01$; on 2-way ethnicity, with $p < 0.01$; and on 3-way ethnicity, with $p < 0.02$. The content baseline is significantly better than our best models, using the same test, with at least $p < 0.0001$.

		Gender		Ethnicity (3-way)		Ethnicity (2-way)	
Training	Model	Acc	F1	Acc	F1	Acc	F1
Twitter	SVM	82.3	82.4	56.5	43.9	66.0	62.7
	CNN	83.1	83.1	**62.0**	42.5	**73.2**	**71.7**
	RNN	**84.3**	**84.3**	60.8	40.9	71.9	69.3
Twitter Auxiliary pre-train	SVM	82.9	83.2	45.9	**44.4**	58.1	60.9
	CNN	83.6	83.5	61.7	40.5	71.7	68.0
	RNN	84.1	84.1	60.2	40.1	70.5	67.3
-	Content	86.2	86.1	81.0	71.6	88.9	88.1

Table 1: Accuracy and F1 on Twitter test data. The best name-based result in each column is bolded.

		Gender		Ethnicity (3-way)		Ethnicity (2-way)	
Training	Model	Acc	F1	Acc	F1	Acc	F1
Auxiliary	SVM	77.7	77.8	44.3	42.5	59.8	60.1
	CNN: (N)	65.8	65.8	53.7	23.3	60.3	44.3
	CNN: (N+S)	64.2	64.4	53.7	23.3	58.7	56.2
	RNN: (N)	63.3	63.2	53.8	24.4	60.7	57.7
	RNN: (N+S)	63.4	63.3	53.7	23.3	60.7	67.3
Twitter	SVM	82.5	82.6	56.1	43.2	64.1	61.1
	CNN: (N)	83.3	83.2	61.9	41.9	70.5	67.2
	CNN: (N+S)	84.0	84.0	**65.9**	45.3	**73.6**	**71.7**
	RNN: (N)	83.5	83.4	60.9	38.4	69.7	63.3
	RNN: (N+S)	83.8	83.8	65.1	44.8	72.5	69.7
Twitter Auxiliary pre-train	SVM	83.6	83.8	49.8	**46.8**	62.1	63.2
	CNN: (N)	83.8	83.8	59.8	46.3	64.8	54.4
	CNN: (N+S)	82.1	82.1	64.2	44.3	71.1	67.5
	RNN: (N)	**84.1**	**84.1**	59.5	45.0	64.2	57.7
	RNN: (N+S)	83.6	83.5	63.2	42.8	70.7	67.3
-	Content	86.7	86.7	79.7	72.8	87.9	87.4

Table 2: Accuracy and F1 on Twitter development data. "N" indicates name alone, "N+S" indicates name and screen name. The best name-based result in each column is bolded.

may not transfer across domains. For the ethnicity auxiliary dataset, our model quickly overfit to the auxiliary data, learning features which did not generalize to the Twitter dataset. With either aggressive regularization or more sophisticated pre-training approaches, we might better utilize the auxiliary data when we have such a limited amount of Twitter data.

We contextualize our results with similar previous work that used other resources and datasets for similar tasks. Rao et al. (2010) reports an accuracy of 72.3% on gender prediction using n-grams and sociolinguistic features in users' tweets. Burger et al. (2011) reports a gender accuracy of 91.8% using user content and profile information, as well as a dev-set accuracy of 89.1% using the user's name field. Our SVM model reproduces the main features from their name model. Jaech and Ostendorf (2015) used character-level morphology induction to learn sub-units from OkCupid usernames, achieving a gender classification accuracy of 74.2% using only a username. Pennacchiotti and Popescu (2011) reports an F1 score of 65.5% on the 2-way ethnicity task, using a combination of features from Twitter profiles, network, and content. In their model that used exclusively profile features, they report an F1 score of 60.9%.[11] Culotta et al. (2015) report F1 scores between 60% and 70% on the 3-way ethnicity comparison using regression and classification approaches, based on whether a user follows specific accounts associated with particular demographics. Although these are not direct comparisons on the same datasets, they demonstrate that our models achieve competitive performance on common demographics tasks while using just names.

5 Limitations

Our methods are limited by the amount of data available per category and the diversity of categories covered. Every dataset we could find was collected in a manner non-representative of Twitter in general, and had a bias towards users in the United States. Such dataset biases may

[11] The authors collected "users who explicitly mention their ethnicity in their profile," implying that profile features could be unfairly predictive.

affect our tool's predictions in ways that are difficult to measure, and should be a consideration in downstream analyses (Wood-Doughty et al., 2017). While the concept of race and ethnicity is a subject of study in social science research (Van den Berghe, 1978), we only consider three of the categories considered by most surveys, due to the very limited available data. To build a tool to adequately classify all widely-used race and ethnicity categories, a great deal of additional data collection and validation is required.

6 Future Work

Despite its limitations, our model improves on previous approaches that require only a single tweet per user by learning a rich representation of the user's names. While the content baseline outperforms our models, our method requires far less data and can be used in settings when it is too slow or costly to download new data. An exploratory experiment found that incorporating our name-based predictions into the content model produced a gender classification accuracy of 91.0%. That this hybrid model improves dramatically over the use of content alone indicates that the two approaches make different kinds of errors and thus could successfully complement each other(Liu and Ruths, 2013). The question of whether different predictors of Twitter user demographics have correlated errors based on user behavior is considered in Wood-Doughty et al. (2017), which offers other suggestions for more robust models.

Further work could also examine how names vary across different domains; while auxiliary government data did not consistently improve performance in our experiments, we expect that username-based features may transfer across different sites (e.g. from Twitter to Reddit) better than content-based features. In the empirical setting of datasets with a single tweet per user, there is still more information we can leverage to infer demographics; Twitter user profiles include optional fields for description, location, and a profile picture.

While extensions may make our methods more accurate or widely applicable, the present work demonstrates that neural character-level models of names can be successfully leveraged for difficult demographic predictions. We hope that these models will make possible low-resource demographic inference in varied domains. Our code and trained classifiers are available as an update to the Demographer package at `http://bitbucket.org/mdredze/demographer`.

7 Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences under grant number 5R01GM114771.

References

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM*, 270.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *ACL*, pages 181–186.

Pierre L Van den Berghe. 1978. Race and ethnicity: a sociobiological perspective. *Ethnic and racial studies*, 1(4):401–411.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on twitter. In *HLT-NAACL*, pages 1010–1019.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *EMNLP*, pages 1301–1309. Association for Computational Linguistics.

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. *ICWSM*, 15:590–593.

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*.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM*, 133:89–96.

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

Mark Dredze, David A. Broniatowski, Michael Smith, and Karen M. Hilyard. 2015. Understanding vaccine refusal: Why we need social media now. *American Journal of Preventive Medicine*.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for twitter: Timing matters. In *NAACL*.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Aaron Jaech and Mari Ostendorf. 2015. What your username says about you. In *EMNLP*.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM*, 13:273–282.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*, pages 188–197.

Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *WWW*, pages 53–54. International World Wide Web Conferences Steering Committee.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. *NLP+ CSS 2016*, page 108.

Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, page 01.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. *ICWSM*, 11:5th.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

Miles Osborne, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin D Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, et al. 2014. Real-time detection, tracking, and monitoring of automatically discovered events in social media. In *ACL*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM.

Edward Velasco, Tumacha Agheneza, Kerstin Denecke, Gan Kirchner, and Tim Eckmanns. 2014. Social media and internet-based data in global systems for public health surveillance: A systematic review. *The Milbank Quarterly*, 92(1):7–33.

Svitlana Volkova and Yoram Bachrach. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12):726–736.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *ACL*, pages 186–196.

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

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.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.