Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter

Zach Wood-Doughty, Praateek Mahajan, Mark Dredze
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218
zach@cs.jhu.edu, pmahaja2@jhu.edu, mdredze@cs.jhu.edu

Abstract

Twitter user accounts include a range of different user types. While many individuals use Twitter, organizations also have Twitter accounts. Identifying opinions and trends from Twitter requires the accurate differentiation of these two groups. Previous work (McCorriston et al., 2015) presented a method for determining if an account was an individual or organization based on account profile and a collection of tweets. We present a method that relies solely on the account profile, allowing for the classification of individuals versus organizations based on a single tweet. Our method obtains accuracies comparable to methods that rely on much more information by leveraging two improvements: a character-based Convolutional Neural Network, and an automatically derived labeled corpus an order of magnitude larger than the previously available dataset. We make both the dataset and the resulting tool available.

1 Introduction

Twitter has been a boon to researchers who study trends in opinions and behaviors at scale (Velasco et al., 2014). Numerous applications from political science (O'Connor et al., 2010), linguistics (Bamman et al., 2014), health (Paul and Dredze, 2017) and the computational social sciences (Schwartz et al., 2013) have utilized Twitter and other social media platforms as a dataset.

Traditional analyses in these fields require the identification of demographic characteristics of individuals. For example, telephone surveys ask demographic panels so as to contextualize the survey results (Kempf and Remington, 2007). As such, analysis of social media platforms has included demographic contextualization (Chen et al., 2015).

However, Twitter and other social media platforms generally do not provide demographic characteristics of users. As such, multiple systems have been developed to automatically infer demographic characteristics of users. Various systems have been shown to perform well at classifying gender (Ciot et al., 2013; Burger et al., 2011), ethnicity (Culotta et al., 2015; Pennacchiotti and Popescu, 2011), and geographic location (Jurgens et al., 2015; Dredze et al., 2013). These classifiers leverage user data to predict these missing demographic attributes; some methods use the tweets written by the user (Al Zamal et al., 2012), while others track who the user follows (Culotta et al., 2015; Jurgens, 2013).

These tools make a central assumption: the account for which demographic inference is performed belongs to an individual. Yet Twitter accounts do not just belong to individuals; the platform is widely used by organizations to represent their interests on the platform, and it makes little sense to infer the gender of an organization. McCorriston et al. (2015) estimated that 9.4% of users on Twitter are brands or organizations. While we address the issue of bots and other types of Twitter accounts in §2.3, we make the simplifying assumption that all accounts on Twitter are either individuals or organizations, and rely on bot detection systems to first filter other types of accounts. When using Twitter data for studies, researchers should not conflate individuals on Twitter with the organizations and brands who use the platform. An analysis of opinions on vaccinations should not treat the official `@CDC` account as a particularly prolific individual, and a study of grassroots political preferences should not use tweets from major political parties as representative of a specific individual's beliefs.

Despite the differences between individual and

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 56–61
NAACL HLT 2018, New Orleans, Louisiana, June 6, 2018.

organizational accounts, most Twitter analyses do not make any such distinction. This is the easiest option and may be a reasonable simplification in some analyses, but conflating the two groups may introduce biases. The only previous readily-available tool for this task is from McCorriston et al. (2015). The authors built a dataset of 19k users annotated as individuals or organizations by crowdsourced annotation. Using a classifier built on metadata features as well as a sample of tweets from the account, they achieved good accuracy at differentiating these account types and released a Python tool. Unfortunately, the solution of McCorriston et al. (2015) poses several problems. First, the tool requires multiple tweets per account. Many analyses focus on single tweets, so obtaining many tweets from the API for each account can be time consuming. Second, while their annotated corpus has high quality labels, it is relatively small. Since only the user labels are released with the annotations, others who wish to train new models on this corpus will suffer over time as accounts are deleted or made private, removing them from consideration. This can be an issue as the models become stale, as behaviors of individuals and organizations on Twitter continue to shift over time (Laroche et al., 2013; Liu et al., 2014; Zhu and Chen, 2015). A larger corpus would maintain its utility for longer, and ideally, the necessary data collection should be as close to automated as possible.

We address these two issues for the task of identifying individuals versus organizations. First, we propose an almost-fully automatic way of constructing a large corpus of annotated individuals and users. Our dataset is almost an order of magnitude larger than that provided by McCorriston et al. (2015). While our data collection uses weak supervision and contains errors, we can achieve comparable accuracy to a method trained on the dataset produced with high-quality annotations by McCorriston et al. (2015). Researchers can use this corpus, or reconstruct a fresh corpus in the future following our approach. Second, we propose a method for classifying individuals versus organizations based on a character-based Convolutional Neural Network (CNN) that examines only a single tweet from a user account, with a focus on the user profile. This simplifies the process of dividing a dataset into individuals and organizations by obviating the need for additional data downloads. By combining our larger corpus and improved model, we obtain results that are comparable to McCorriston et al. (2015).

2 Data

Our goal was the construction of a large set of Twitter accounts annotated as individual or organization. Rather than rely on manual labeling of accounts, we seek an automated method based on weak supervision (Li et al., 2014) for the discovery and labeling of these accounts. We describe our process in this section, and evaluate the efficacy of our resulting dataset by evaluating models trained on this corpus.

2.1 Twitter Lists

Twitter users can create "lists," collections of Twitter accounts organized by topic. Examples of lists include "social-justice organizations" or "volleyball teammates." Lists are useful ways for crowdsourcing the identified and organization of Twitter accounts.

We identified Twitter lists that predominantly contained either organizations or individuals. We used a search engine to find user-generated lists which included key terms such as "businesses" or "companies" to identify lists of organization accounts. For each list that we verified as likely containing organizations, we downloaded the Twitter accounts that were members of the list and labeled them organizations. We repeated this process for individuals by searching for terms such as "friends" or "family."

Using this approach to gather about 250 lists, we collected 19k accounts labeled as individuals and 28k accounts labeled as organizations. After data collection was complete, we randomly sampled 100 organizations and 100 individuals for verification, and found 98% were labeled correctly.

2.2 LinkedIn

We identified individuals on Twitter through the presence of a link to a LinkedIn profile page in the users' Twitter profile. We examined the `user.url` field for links with the domain `linkedin.com` or `lnkd.in`. We examined all tweets collected from Twitter's 1% feed in 2017, about 3 billion tweets. We then extracted the set of unique authors of these tweets, yielding a corpus

of 161k users we believe to be individuals. After finishing data collection, we randomly sampled 100 of these accounts and found that all were correctly labeled.

In total, these two methods produced a list of 180k individuals and 28k organizations.

2.3 Limitations

Our work makes the simplifying assumption that all accounts are either individuals or organizations, and ignores other possible types of Twitter accounts. We assume that accounts are first processed by bot detection systems to identify them as either "human and non-human" users (Dickerson et al., 2014), where the non-human users can be subdivided into "spambots, paybots, or influence bots" (Subrahmanian et al., 2016). In this work, we treat these bot categories as orthogonal – that is, a spambot or influence bot may *impersonate* an individual or an organization, but our tool only considers this latter designation. This simplifying assumption may be reasonable given the data we consider. In the dataset constructed by McCorriston et al. (2015), human annotators were only allowed to code a random selection of Twitter users as either individuals or organizations, and yet 90.7% of the accounts had a unanimous labeling across three annotators, with a inter-annotator Cohen's κ=0.95. Twitter bots who cannot be labeled as individuals or organizations may exist, but we expect they are rare. Further research should consider the correlation between our tool's predictions and the predictions made by systems such as BotOrNot (Davis et al., 2016) or SentiBot (Dickerson et al., 2014). Future work could improve our tool by incorporating features used by these bot classifiers, though many such features cannot be computed when using only one tweet per user.

In the lists and LinkedIn data we collected, we found that these methods identified accounts that agreed with the inferred label with high probability. However, some labels may be blatantly wrong and others may be ambiguous in the eyes of human annotators. Twitter lists are generated and named by users, and may have misleading titles or contain erroneous accounts. Similarly, some organizational accounts may link their Twitter profile to a `linkedin.com` page, which would cause us to incorrectly label their account in our dataset.

A second drawback of our training data is that it is not drawn from a representative sample of the Twitter user population. Accounts which are added into other users' lists are likely more popular than a randomly-selected account, and individuals who link their Twitter account to a LinkedIn page likely present a more professional appearance in their profile or tweets. This may bias our classifier to misjudge less popular organizational users or the accounts of individuals who do not use Twitter professionally.

We evaluate the impact of our data limitations by using this corpus as a training set for classifications on a high-quality test set. We leave the considerations of Twitter bots for future work.

3 Methods

We present three models for the task of classifying users as either organizations or individuals: a baseline method and two new methods that require only a single tweet per user.

3.1 Humanizr

McCorriston et al. (2015) proposed a method (named Humanizr) for classifying individuals and organizations based on features extracted from the profile and tweet history of Twitter users. This method requires the downloading of multiple tweets for each account for classification. The extracted features are then used by an SVM to learn a binary classifier. We used their released code to train their models on a March 2018 download of their dataset, for which we successfully retrieved 86% of the users.

3.2 Ngram Model

Knowles et al. (2016) developed a model for gender classification of Twitter users based on an linear model trained on character n-gram features from users' names. They found that their model outperformed several available baselines for gender prediction. Furthermore, since the model considered the username, it required only a single tweet from a user to make a prediction. We extend their n-gram feature selection by incorporating new name-based features which we expect to be indicative of the organization versus individual task, such as the occurrence of capitalization and numeric characters. We combine these name-based features with the profile-based features described below.

	Balanced	Full
Majority	50.0	90.0
Humanizr	**89.6**	**94.8**
N-gram	85.2	93.8
CNN	84.5	93.4

(a) Results from training on the data released by McCorriston et al. (2015).

	Balanced	Full
Majority	50.0	90.0
Humanizr	-	-
N-gram	84.0	94.1
CNN	**85.8**	**94.6**

(b) Results from training on our collected data. Humanizr was not evaluated due to data constraints.

Figure 1: Experimental results. In both experiments, the test sets are 20% of the data released by McCorriston et al. (2015).

3.3 Convolutional Neural Network

We use a character-based CNN to learn a representation of a Twitter user's name. After some initial experiments, we used a simple stack of two convolutional layers of 256 filters of width 3. The name representation learned by the CNN is concatenated with the profile-features described below, and this combined vector is passed through two fully-connected layers to produce a distribution over the labels.

3.4 Profile-based features

Both our Ngram and CNN models incorporate features extracted from the user fields contained in the metadata of a single tweet object. Some of these features – the ratio of followers to friends, verification status, and the number of tweets – were used in previous work (McCorriston et al., 2015). We also introduce new features, such as the presence of personal pronouns (e.g. "my" vs. "our") and the use of repetitive punctuation (e.g. "!!") in users' descriptions. A complete list of our profile features is included in the released code. For all continuous features (e.g. follower to friends ratio), we normalize them to take values between -1 and 1 using a piecewise linear function constructed from their deciles.

3.5 Parameter Estimation

For both the n-gram and CNN models, we used the held-out development set for hyper-parameter tuning. For the n-gram model, we considered hinge or perceptron loss functions, and L1 or L2 regularization, using the implementations from sklearn (Pedregosa et al., 2011). For the CNN, experimentation led us to use a SGD optimizer with learning rate 0.5, a character embedding of 256, and dropout rate of 0.2, using implementations in Tensorflow (Abadi et al., 2016). We train for up to 200 epochs, using the dev set for early stopping.

4 Evaluation

We ran two experiments, each focused on one major question. First, how well do our proposed models perform on this task, when using only a single tweet per user, compared to the Humanizr method? Second, how useful is the dataset we created for training models to discriminate between organizations and individuals?

To answer the first, we apply our two methods to the dataset collected by McCorriston et al. (2015) and compare against their method. We take data for the 17k users we could scrape and split them into train, dev, and test sets. We do this for two experimental settings: a 'balanced' setting in which we subsample individuals so that each split has an equal number of individuals and organizations, and a 'full' setting in which we use the ratio of individuals to organizations (approximately 8:1) that occurs in the scraped data. Empirically, we found that a training set ratio of 7:1 individuals to organizations improved dev performance in the class-imbalanced, 'full' experimental setting.

To answer the second question, we use the dataset we collected as training data, but use the McCorriston et al. (2015) data for dev and test sets. This examines whether the features learned from Twitter users in our noisy and cheap dataset are useful for classifying users in a high-quality and expensive dataset. In this experiment, we did not evaluate the Humanizr method due to the cost associated with downloading 200 tweets per user for 180k users. We again considered both balanced and full experimental settings.

For each experimental setting, we use 20% of the McCorriston et al. (2015) data for the dev and test sets, either class-balanced or not. To highlight the difference between the balanced and full settings, we include the proportion of the majority label as a baseline classification accuracy.

4.1 Results

Table 1 (a) shows the results for the first experiment. While the Humanizr method slightly outperforms both our n-gram and CNN models, it does so while using significantly more data per user. The Humanizr method's test accuracy on our splits was slightly lower than the five-fold cross validation accuracy reported in McCorriston et al. (2015). This may be because we were unable to download 14% of the users in the original dataset or because we did not retune their hyper-parameters to the tweet data from 2018.

Table 1 (b) shows the result for the second experiment, evaluating our models trained on our collected dataset. The CNN improves considerably, almost matching the performance of Humanizr. In fact, in the full setting, the difference between the two is not statistically significant.[1] This provides strong evidence that our dataset, while cheaply collected with noisy labels, is valuable for classifying organizations and individuals on a random sample of Twitter.

While the n-gram model slightly outperformed the CNN in the first experiment, the trend was reversed in the second. This may be because the smaller dataset in the first experiment was sufficient for our hand-engineered n-gram features, but not large enough for the CNN models to learn robust character-level features from data alone.

Together, these two experiments demonstrate that a method which requires just a single tweet per user can be trained on cheaply-gathered data to classify organizations on Twitter, and perform comparably to a tool trained on high-quality data with hundreds of tweets per user. Our method makes it possible to classify organizations in an analysis of billions of tweets without having to download significant additional data per user. Our method also makes possible analyses in a streaming setting in which their decisions must be made in real-time without additional data collection.

Future work should see whether our tool's predictions are correlated with the predictions of bot-detection systems, and whether our model could be used to predict bot or other non-human account types. We could also incorporate the content features from Humanizr with our name and profile features we introduce. Another avenue for future work is to consider whether we can control for any biases in our weakly-supervised dataset to produce better predictions on the ground-truth data. As it is often easier to collect a large amount of noisy data than a small amount of gold-standard data, such an approach could be widely applicable to studies of Twitter users' emotions and personalities.

We release the account-type labels and the Twitter userids for our training dataset, as well as our code for our feature extraction and experiments. We also provide a pre-trained model for classification of Twitter accounts. The code, data, and models are available as an extension to the Demographer tool at `http://bitbucket.org/mdredze/demographer`.

[1] p=0.36 when using a two-proportion z-test. For the balanced setting, Humanizr's 89.6% is significantly better than the best CNN's 85.8%, with p=0.014 using the same test.

5 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.

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

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

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

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *EMNLP*, pages 1136–1145.

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

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *WWW*, pages 273–274. International World Wide Web Conferences Steering Committee.

John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *ASONAM*, pages 620–627. IEEE.

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)*.

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.

Angela M Kempf and Patrick L Remington. 2007. New challenges for telephone survey research in the twenty-first century. *Annu. Rev. Public Health*, 28:113–126.

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

Michel Laroche, Mohammad Reza Habibi, and Marie-Odile Richard. 2013. To be or not to be in social media: How brand loyalty is affected by social media? *International Journal of Information Management*, 33(1):76 – 82.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. In *ACL*, volume 1, pages 165–174.

Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. 2014. The tweets they are a-changin: Evolution of twitter users and behavior. In *ICWSM*, volume 30, pages 5–314.

James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *ICWSM*, pages 650–653.

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.

Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

V. S. Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, Rand Waltzman, Andrew Stevens, Alexander Dekhtyar, Shuyang Gao, Tad Hogg, Farshad Kooti, Yan Liu, Onur Varol, Prashant Shiralkar, V. G. Vinod Vydiswaran, Qiaozhu Mei, and Tim Huang. 2016. The DARPA twitter bot challenge. *CoRR*, abs/1601.05140.

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.

Yu-Qian Zhu and Houn-Gee Chen. 2015. Social media and human need satisfaction: Implications for social media marketing. *Business horizons*, 58(3):335–345.