Simple Queries as Distant Labels for Predicting Gender on Twitter

Chris Emmery[1,2] and **Grzegorz Chrupała**[1] and **Walter Daelemans**[2]
[1]TiCC, Tilburg University, 5000 LE Tilburg, The Netherlands
[2]CLiPS, University of Antwerp, Prinsstraat 13, B-2000 Antwerpen, Belgium
{c.d.emmery, g.a.chrupala}@uvt.nl
walter.daelemans@uantwerpen.be

Abstract

The majority of research on extracting missing user attributes from social media profiles use costly hand-annotated labels for supervised learning. Distantly supervised methods exist, although these generally rely on knowledge gathered using external sources. This paper demonstrates the effectiveness of gathering distant labels for self-reported gender on Twitter using simple queries. We confirm the reliability of this query heuristic by comparing with manual annotation. Moreover, using these labels for distant supervision, we demonstrate competitive model performance on the same data as models trained on manual annotations. As such, we offer a cheap, extensible, and fast alternative that can be employed beyond the task of gender classification.

1 Introduction

The popularity of social media that rely on rich self-representation of users (e.g. Facebook, LinkedIn) make them a valuable resource for conducting research based on demographic information. However, the volume of personal information users provide on such platforms is generally restricted to their personal connections only, and therefore off-limits for scientific research. Twitter, on the other hand, allows only a restricted amount of structured personal information by design. As a result, their users tend to connect with people outside of their social circle more frequently, making many profiles and communication publicly accessible. A wide variety of research has long picked up on the interesting characteristics of this microblogging service, which is well facilitated by the Twitter REST API.

The applied Natural Language Processing (NLP) domain of author profiling aims to infer unknown user attributes, and is therefore broadly used to compensate for the lack thereof on Twitter. While previous research has already proven to be quite effective at this task using predictive models trained on manual annotations, the process of hand-labelling profiles is costly. Even for the ostensibly straight-forward task of annotating gender, a large portion of Twitter users purposefully avoids providing simple indicators such as real names or profile photos including a face. Consequently, this forces annotators to either dive deep into the user's timeline in search for linguistic cues, or to make decisions based on some personal interpretation, for which they have shown to often incorrectly apply stereotypical biases (Nguyen et al., 2014; Flekova et al., 2016).

We show that running a small collection of ad-hoc queries for self-reports of gender once ("I'm a male, female, man, woman" etc.) — provides distant labels for 6,610 profiles with high confidence in one week worth of data. Employing these for distant supervision, we demonstrate them to be an accurate signal for gender classification, and form a reliable, cheap method that has competitive performance with models trained on costly human-labelled profiles. Our contributions are as follows:

- We demonstrate a simple, extensible method for gathering self-reports on Twitter, that competes with expensive manual annotation.
- We publish the IDs, manual annotations, as well as the distant labels for 6.6K Twitter profiles, spanning 16.8M tweets.

The data, labels, and our code to collect more data and reproduce the experiments is made available open-source at https://github.com/cmry/simple-queries.

2 Related Work

Author profiling applies machine learning to linguistic features within a piece of writing to make inferences regarding its author. The ability to make such inferences was first discussed for gender by Koppel et al. (2002), and initially applied to blogs (Argamon et al., 2007; Rosenthal and McKeown, 2011; Nguyen et al., 2011). Later, the work extended to social media — encompassing a wide variety of attributes such as gender, age, personality, location, education, income, religion, and political polarity (Eisenstein et al., 2011; Alowibdi et al., 2013; Volkova et al., 2014; Plank and Hovy, 2015; Volkova and Bachrach, 2016). Apart from relevancy in marketing, security and forensics, author profiling has shown to positively influence several text classification tasks (Hovy, 2015).

Gender profiling research on Twitter generally takes a data-driven, open-vocabulary approach using bag of words, or bag of n-gram features (Alowibdi et al., 2013; Ciot et al., 2013; Verhoeven et al., 2016), applying supervised classification using manually annotated profiles. However, distant supervision has as of yet only looked at non-textual cues for this task, unlike for example age, personality, and mental health (e.g. Al Zamal et al., 2012; Plank and Hovy, 2015; Coppersmith et al., 2015). For gender, Burger et al. (2011) and Li et al. (2014) collect links to external profiles, whereas Al Zamal et al. (2012) and Li et al. (2015) use a list with gender-associated names. Both of these approaches rely on continuous monitoring of streaming data, and utilize indicators that are typically easy cues for annotators, thereby omitting profiles that would be costly to annotate. In contrast, our method only has to be repeated once a week, and includes a different set of users where sampling is not influenced by external resources.

3 Data Collection

To empirically compare distant labels (i.e. obtained using heuristics) with manual annotations, we require both data containing self-reports, and corpora with hand-labelled Twitter profiles for comparison.

Distant Labels The profiles in our corpus were collected on March 6th, 2017 — using the Twitter Search API[1] to query for messages self-reporting gender: e.g. {`I' / I a`}`m a` {`man, woman, male, female, boy, girl, guy, dude, gal`}. For each retrieved tweet, the timeline of the associated author was collected (up to 3,200 tweets) between Match 6th and 8th. Note that the maximum retrieval history for the Search API is limited to tweets from the past week. Hence, our set of queries collected 19,307 profiles spanning results for one week only.

This method has some inherent advantages in addition to the ones mentioned in Section 2: it guarantees to a large extent that the profiles gathered are primarily English (95% of all associated tweets), collects data from active users (average of 2,500 tweets per timeline), and generally avoids bots, or other spam profiles (0,2%[2] of all profiles). Finally, with gender profiling being considered a binary male/female classification task for much of the previous research and corpora, it also prevents including users that might not identify with the binary framework in which gender is typically cast.[3]

filter	N hand	F	F+R
`none`	1,456	.806	.806
`rt`	1,109	.873	.887
`rt + "`	1,059	.882	.896
`rt + :`	1,091	.887	.891
`rt + " + :`	1,045	.885	**.900**

Table 1: Several filter rules applied to the distant labels (effectively removing those matching the rules), their impact on both data reduction (N hand-labelled) and agreement increase. Agreement is specified for: only applying these filters (F), and in combination with the rules from Table 2 (F+R), and reflects the amount of correct distant labels compared to the manual labels.

Manual Evaluation To evaluate the accuracy of our distant labels, a random sub-sample was manually labelled for gender by two annotators using a full profile view ($\kappa = 0.78$), resulting in 1,456 agreed on labels. Based on the initial results (see Table 1), several rules were constructed to filter (thereby removing) any profiles the query tweet matched to. First, we observed that many tweets (31%) contained `rt` — indicating a retweet. Similar to tweets containing quotes (5%), or colons

[1] https://dev.twitter.com/rest/public/search

[2] Bots were identified during annotation.

[3] Accordingly, this method could be applied in future research tackling this long-standing issue by collecting and using self-reported non-binary representations of gender.

Location	Rule set
anywhere	according to, deep down
before query	feel like, where, (as) if, hoping, assume(s/d) (that), think, expect (that), then, (that) means, implying, guess, think(s), tells me

Table 2: Rules applied to the distant labels to flip the assumed gender. Their location can be *anywhere* in the tweet, or right *before* the *query* (e.g. "Sometimes I think I'm a girl").

	Volkova	Plank	Query
users	4,620	1,391	6,610
tweets	12,226,859	3,568,265	16,788,612
female	32,367	10,613	61,736
male	26,708	6,739	32,900
train	47,298	13,827	75,918
test	11,777	3,525	18,718

Table 3: Various metrics of the Twitter corpora annotated with gender used in this research. The train and test sizes reflect the amount of batches of 200 tweets.

(2%), these are generally not self-reports (e.g. "random guy: I'm a man..."), and were therefore removed. Overall, the filters increased agreement with our manual annotations, simultaneously causing a decrease to 6,610 profiles. This method however ensures a high accuracy of the distant labels, which should outweigh the amount of data.

In addition to these filters, several rules were constructed to deal with linguistic cues that make it highly likely for the gender to be the opposite of the literal report (see Table 2) — thus indicating the label should be flipped. Examples include "according to the Internet, I'm a girl", and "Don't just assume I'm a guy". For a detailed overview of their effect on the overall agreement, see F+R in Table 1. The ad-hoc list presented here improved agreement about .015. Note that despite being constructed by manual inspection of the mismatches between annotations and the distant labels, our filters, rules, and even the initial query can be extended with some creativity.

Preparation To compare our distant labels to annotated alternatives, we include Volkova et al. (2014)'s crowd-sourced corpus, and the manually labelled corpus by Plank and Hovy (2015). Henceforth, these external corpora will be referred to as Volkova and Plank respectively. The timelines of their provided user IDs where gathered between April 1st and 7th 2017 (see Table 3 for further details on their sizes).

The timelines for all corpora—including our Query corpus—were divided in batches of 200 tweets, as most related work follows this setup. Afterwards, each batch is provided with either a distant, or manual label, depending on the set of origin. This implies that users with less than 200 tweets were excluded, as well as any consecutive tweets that would not exactly fit into a batch of 200. The corpora were divided between a (gender stratified) train and a test set by user ID. This guarantees that there is no bleed of batches from any user between any of the splits (refer to Table 3 for the final split sizes). Other than tokenisation using `spaCy` (Honnibal and Johnson, 2015), no special preprocessing steps were taken. We removed primarily non-English batches using `langdetect`[4] (Shuyo, 2010), as well as the original query tweets containing self-reports. The latter was done to avoid our queries being most characteristic for some batches.

4 Experiment

For document classification, `fastText`[5] (Joulin et al., 2016) was employed; a simple linear model with one hidden embedding layer that learns sentence representations using bag of words or n-gram input, producing a probability distribution over the given classes using the softmax function. It therefore follows the same architecture as the continuous bag of words model from Mikolov et al. (2013), replacing the middle word with a label. Joulin et al. (2016) demonstrate the model performs well on both sentiment and tag prediction tasks, significantly speeding up training and test time compared to several recent models.

Gender predictions were made using a typical set of n-gram features as input; token uni-grams and bi-grams, and character tri-grams. We incorporate only those grams that occur more than three times during training. As the corpora are quite small, we use embeddings with only 30 dimensions, a learning rate of 0.1, and a bucket size of 1M. All models are trained for 10 epochs. Given that `fastText` uses Hogwild (Recht et al., 2011)

[4] https://github.com/Mimino666/langdetect

[5] https://github.com/facebookresearch/fastText

				Train		
		Majority	**Lexicon**	**Volkova**	**Plank**	**Query**
Test	Volkova	.556	.796	**.822** (0.001)	.701 (0.007)	.771 (0.007)
	Plank	.659	.740	**.741** (0.005)	.723 (0.003)	.724 (0.009)
	Query	.674	.668	.730 (0.007)	.689 (0.005)	**.756** (0.002)
	Average	.630	.735	.764	.704	.750

Table 4: Individual accuracy scores and averages for majority baseline (Majority), the lexicon of Sap et al. (2014), and the three models (trained on Volkova, Plank, and our dataset respectively) evaluated on the test set for each corpus. Standard deviation is reported after repeating the same experiment 20 times.

for parallelising Stochastic Gradient Descent, randomness in the vector representations cannot be controlled using a seed. To estimate the standard deviation in the results, we ran each experiment 20 times. To evaluate how our distantly supervised model compares to using manual annotations, we trained all models in this same configuration for all three corpora. Each model was then evaluated on the test set for each corpus.

5 Results

Table 4 shows accuracy scores for this 3x3 experimental design, as well as a majority baseline score (always predicting female), and an average over the three test sets for each model. We closely reproduced the results from Volkova and Bachrach (2016); despite the difference in user[6] and tweet samples, exact split order, and their use of more features including style and part-of-speech tags, our performance approaches their reported .84 accuracy score. Plank and Hovy (2015) do not provide classification results for gender on their data. For comparison to state-of-the-art gender classification for English, the lexicon of Sap et al. (2014) is included in the results. Their work also compares with Volkova et al. (2014), and reports a higher score (.90) for their random sample setup than reproduced in our batch evaluation (.80).

Despite the fact that the model trained on the Volkova corpus performs best on both annotated corpora (Volkova and Plank), the difference is fairly small compared to our distantly supervised model — the latter of which somewhat expectedly performs best on its associated test set. On average, the Query and Volkova trained models only differ .014 in accuracy score, and the Query model outperforms the lexicon approach by .015. However, the more significant comparison is the out of sample performance for these two models and the lexicon model on the Plank test set. Here, results are comparable between Query and Volkova, with a .017 difference, and higher standard deviation. However, here the lexicon approach outperforms the Query model with .016. Not only does this show our distant labels to be be comparable with hand labels, our models also seems to yield favourable performance over state of the art.

[6]We could only retrieve 4,620 of the reported 4,998.

6 Conclusion

We use simple queries for self-reports to train a gender classifier for Twitter that has competitive performance to those trained on costly hand-annotated labels — showing minimal differences. These should be considered in light of the manual effort put into gathering the annotations, however. Labelling Twitter users with our set of queries yields up to 45,000 hits per 15 minutes (API rate limits considered), and therefore finishes in several minutes. Retrieving the timelines for the initial 19,307 users took roughly 21 hours. Including preprocessing (3 hours) and running `fastText` (a few minutes) the entire pipeline is encouragingly cheap, even considering time, and can feasibly be repeated on a weekly basis.

Hence, through manual analysis, as well as experimental evidence, we demonstrate our distantly supervised method to be a reliable and cheap alternative. Moreover, we pose several ways of improving this method by extending the queries, and further fine-tuning the applied filters and rules for a correct interpretation of the reports. By altering the queries to match other types of self-reports, it offers the possibility of quickly exploring its effectiveness for inferring other user attributes with little effort. We hope to facilitate this for the research community by providing our implementation. Our further work will focus on intelligently expanding the queries and evaluating this method on a larger scale with more attributes.

References

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

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*. IEEE, volume 1, pages 365–369.

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

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1301–1309.

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

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. *NAACL HLT 2015* page 1.

Jacob Eisenstein, Noah A Smith, and Eric P Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 1365–1374.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preotiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*. pages 843–854.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1373–1378.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *ACL*. pages 752–762.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* .

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4):401–412.

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

Jiwei Li, Alan Ritter, and Dan Jurafsky. 2015. Learning multi-faceted representations of individuals from heterogeneous evidence using neural networks. *arXiv preprint arXiv:1510.05198* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, pages 115–123.

Dong-Phuong Nguyen, RB Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and FMG de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Association for Computational Linguistics.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 92–98.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*. pages 693–701.

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*. Association for Computational Linguistics, pages 763–772.

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

Nakatani Shuyo. 2010. Language detection library for java. *Retrieved Jul* 7:2016.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, ELRA, Portorož, Slovenia.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *ACL (1)*. pages 186–196.