Observational Comparison of Geo-tagged and Randomly-drawn Tweets

Tom Lippincott and **Annabelle Carrell**
Johns Hopkins University
tom@cs.jhu.edu, belle.carrell@gmail.com

Abstract

Twitter is a ubiquitous source of micro-blog social media data, providing the academic, industrial, and public sectors real-time access to actionable information. A particularly attractive property of some tweets is *geo-tagging*, where a user account has opted-in to attaching their current location to each message. Unfortunately (from a researcher's perspective) only a fraction of Twitter accounts agree to this, and these accounts are likely to have systematic diffences with the general population. This work is an exploratory study of these differences across the full range of Twitter content, and complements previous studies that focus on the English-language subset. Additionally, we compare methods for querying users by self-identified properties, finding that the constrained semantics of the "description" field provides cleaner, higher-volume results than more complex regular expressions.

1 Motivation

Twitter users can opt-in to include their current geographic location with their tweets. This fine-grained information has been used for a number of down-stream tasks, including bot and spam account detection ((Guo and Chen, 2014)), demographic analysis ((Malik et al., 2015), (Pavalanathan and Eisenstein, 2015)), and enhancing situational awareness for disaster or public health crises ((Amirkhanyan and Meinel, 2016)).

As many of these studies note, there are a number of reasons to be suspicious of geo-tagged tweets as a direct source of realistic communications between people. Popular media has raised public awareness of the dangers in sharing one's location, while for a non-human user (e.g. a business, pseudonymous personality, government entity) this may be exactly the information intended for dissemination. More specific factors like country, culture, and technology further complicate the relationship between geo-tagged accounts and the general user base.(Sunghwan Mac Kim and Paris, 2016; Karbasian et al., 2018)

2 Previous studies

A number of prior work has investigated how Twitter users, and subsets thereof, relate to more general populations. (Malik et al., 2015) collate two months of geo-tagged tweets originating in the United States with county-level demographic data, and determine that geo-tagged users differ from the population in familiar ways (higher proportions of urban, younger, higher-income users) and a few less-intuitive ways (higher proportions of Hispanic/Latino and Black users). (Sloan et al., 2015), (Sloan, 2017) combined UK government and targeted surveys, human validation, and information from user descriptions to compare Twitter and general population distributions over age and occupation, reporting significant differences between both the data sets and the quality of classifiers. (Pavalanathan and Eisenstein, 2015) compared aggregate properties of tweets whose location was determined from geo-tagging with those determined from the free-form user "location" field. They focused on the 10 large urban centers in the US, and found significant variation in age and gender demographics. They note that such differences, which are correlated with linguistic properties and classification difficulty for automatic geo-tagging, and the higher activity of geo-tagged users, can produce inflated accuracies as an evaluation set. These studies focused on English-language data, and regions in either the United States or United Kingdom: this study expands attention to previously-unstudied languages and geographies.

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 50–55
NAACL HLT 2018, New Orleans, Louisiana, June 6, 2018.

3 Methods

We used Twitter's streaming API to collect a *geo-tagged* (GT) data set of all geo-tagged content from the final week of November 2017, and a *non-geo-tagged* (NGT) data set of the 1% uniform random sample from the same time period, *minus* geo-tagged content. We then indexed the tweet and user JSON objects in ElasticSearch(Gormley and Tong, 2015) to facilitate comparisons between the two data sets. After examining several high-level properties, we chose *language*, *hash tag*, *user mention* and *time zone* as the most well-populated categorical fields to focus on.[1]

Following the work of (Beller et al., 2014) we extracted user self-identification in tweets based on the case-insensitive regular expression "I('m|am) an? (\S+)", limiting our results to the same 33 roles considered in that study. We also target the same set of roles by simply querying for users whose "description" field contains the role. The authors examined 20 randomly-chosen hits for each combination of role and methodology to determine precision, shown in Figure 1. We consider pattern matches on "retweets" to be false positives. Interestingly, despite its relative simplicity, the description queries are almost universally more precise, while also pulling back orders of magnitude more results. We therefore use it as our source for this demographic information, and perform the same comparisons for role-distributions as for other categorical fields. Note that our focus on precision is partly due to our focus on building high-quality training data sets, and partly due to the difficulty of measuring recall, particularly for low-frequency roles. We leave this for future work.

To compare distributions over discrete outcomes (e.g. GT versus NGT language use, role occurrence, etc) we calculate the Jensen-Shannon divergence (JSD)(Grosse et al., 2002), a symmetric variant of the Kullbeck-Liebler divergence.

Finally, we compared the same discrete features conditioned on *language*, with the hypothesis that possible causes like spam and commercial content may be particularly focused on particular communities for which language is a reasonable proxy. To explore whether different axes of GT-NGT variation (e.g. hashtags, roles) behave across different language communities, we calculate the Spearman rank correlation coefficient(Hollander et al., 2013) over the JSD values.

[1]Twitter's terms of service prevent distribution of the underlying data, but we make the fine-grained counts available as pickled query results from ElasticSearch at www.cs.jhu.edu/~tom/naacl18_PEOPLE_ES_query.pkl.gz

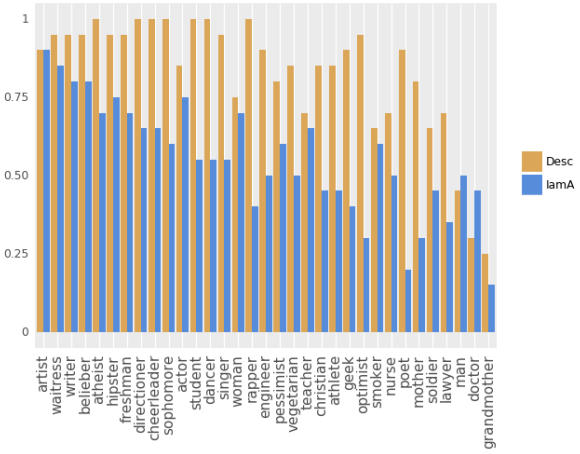

Figure 1: Precision of roles extracted via the "IamA" pattern versus the "description" field

4 Results

4.1 Macro-level comparisons

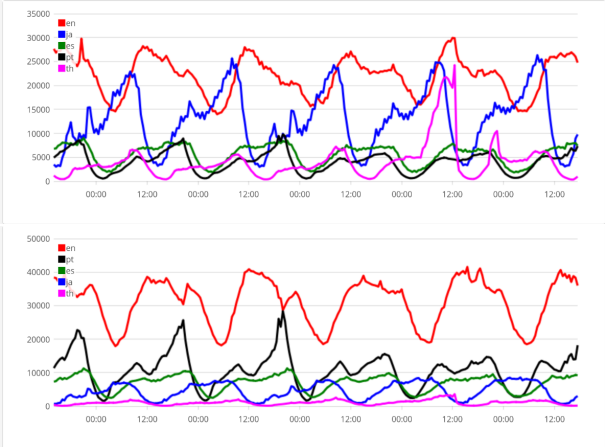

Figure 2: Comparison of GT and NGT tweet volume for several languages over one week

Figure 2 compares GT and NGT tweet volume over time in several high-frequency languages. The expected diurnal pattern from Twitter's overall language distribution is accentuated by the GT skew towards English and Portuguese, with large populations in the Americas. The sharp spike in NGT for Thai is due to a high-profile contestant in the Miss Universe competition.

The number of tweets collected in the GT and NGT data sets is of similar scale (28.5m and 23m,

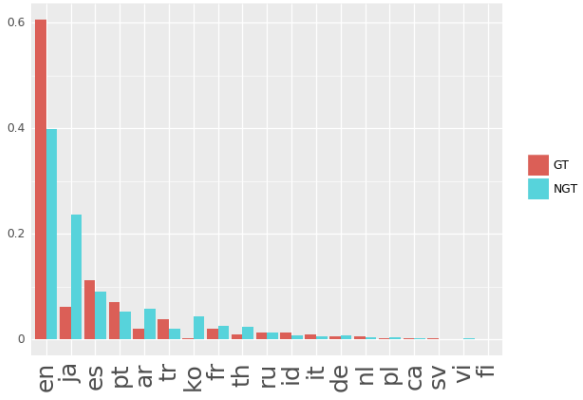

Figure 3: User languages

respectively) but GT users tweet at over triple the rate (8.4 and 2.5 average per user, respectively).[2] Additionally, GT accounts tend to be about twice the age of NGT accounts (Dec. 2012 and Feb. 2015 average creation dates, respectively), and 1% of GT users are verified, compared to NGT at 0.5%.[3] Table 1 shows aggregate information related to how users in each data set participate in Twitter's community structure on average.

Data	Friends	Followers	Favorites
GT	670	2096	4912
NGT	601	1569	4408

Table 1: Average counts of user behavior

Note that, in all of these dimensions, the GT users appear to be more active and engaged with Twitter's structure. How this behavior is attributable to self-selectiveness of individuals, the nature of institutional and spam accounts, or other causes is an open question.

Figure 3 compares user distributions over languages. Among the most common languages, Japanese, Arabic, Thai, and particularly Korean-language accounts have low proportions of geotagging, while Spanish, Portuguese, and particularly English and Turkish have high proportions.

The time zone comparison reflects similar trends, and also allows zeroing in on some specific locales, like Irkutsk, Baghdad, and Paris. It would be useful to determine the various ways in which the *time zone* field can be set, perhaps in tandem with source information (device, app), to better understand this data.

4.2 Hash tags and user mentions

Figures 4 and 5 compare counts of the most-frequent hash tags and user mentions, respectively. Hash tags are dominated by discussion of the Miss Universe competition, particularly from Thailand. Discounting such one-off events, the majority of tags are English-language and related to potential employment, with general values like *job*, *CareerArc*, *hiring*, and industries like *Hospitality*, *HealthCare*, *CustomerService*. These are almost universally geo-tagged, supporting the hypothesis that institutional accounts are a likely source for much of the geo-tagged content. Not visible in the figure, tags relating to various cryptocurrencies tend to *not* be geo-tagged, perhaps reflecting cultural and technological aspects of that demographic.

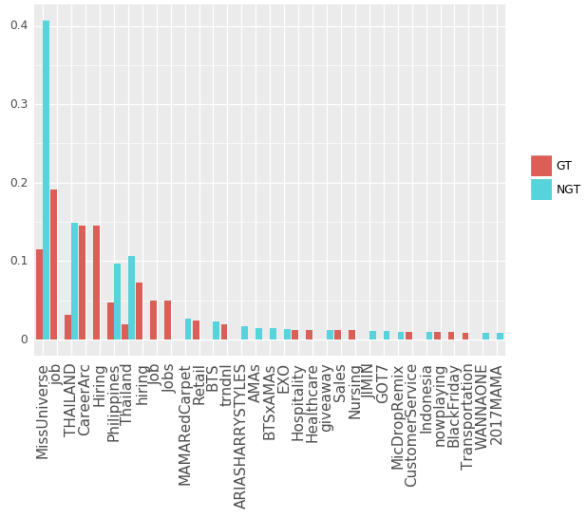

Figure 4: Comparison of most-frequently-used hashtags

Geo-tagged users most frequently mention accounts that are institutional (publicly-traded companies, news organizations, sports franchises) with the notable outliers of accounts associated with Donald Trump, while NGT users are more likely to mention pop stars.[4] Most institutional accounts are only mentioned by GT users, likely self-referentially (e.g. *StarbucksTR*, *NissanUSA*)

[2]We thank a reviewer for pointing out a methodological problem with the original comparison: however, we performed the same comparison of between the full account histories of GT and NGT users from a large window in the 1% sample, and found the same proportion.

[3]Twitter recognizes accounts that are "maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas"

[4]This is likely biased by services that transfer messages from other social media service in e.g. Asia, which appear to not include geotagging

and more for broadcasting information than active engagement. *FoxNews* is an outlier in this respect, as NGT users often address it directly.

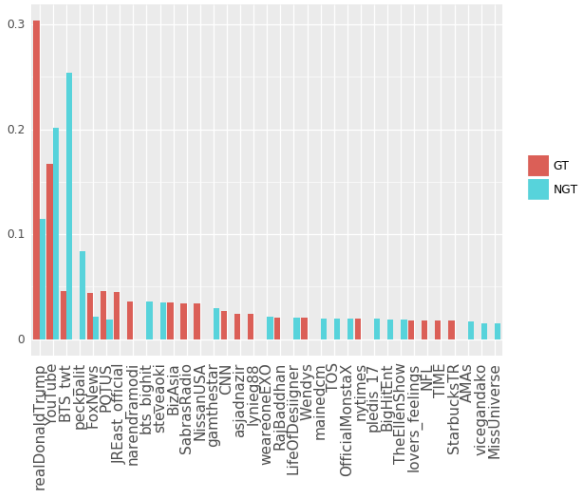

Figure 5: Comparison of most-frequently-mentioned user accounts

4.3 Self-identification

Figure 6 compares relative frequency of each role in the GT and NGT data sets, which have a high Spearman correlation of 0.944. Roles focusing on religion (*Christian*, *atheist*) and musical fandom (*Belieber*, *Directioner*) have a strong preference against geo-tagging, while roles involving performance (*singer*, *actor*, *athlete*, *cheerleader*) seem more inclined to publicize location.

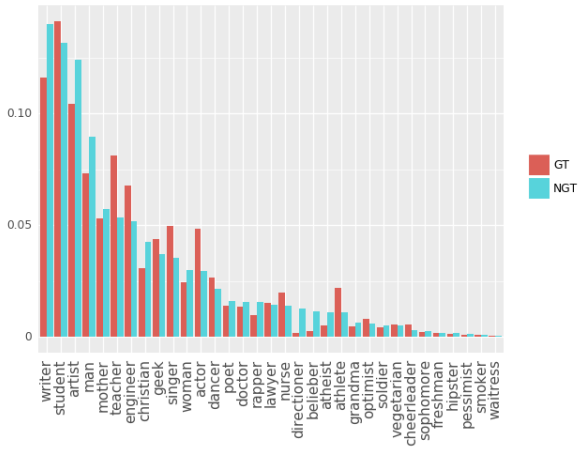

Figure 6: Comparison of role frequencies between the data sets, extracted from user descriptions

4.4 Variation by language

Figure 7 plots JSD divergence between GT and NGT distributions over several discrete spaces.

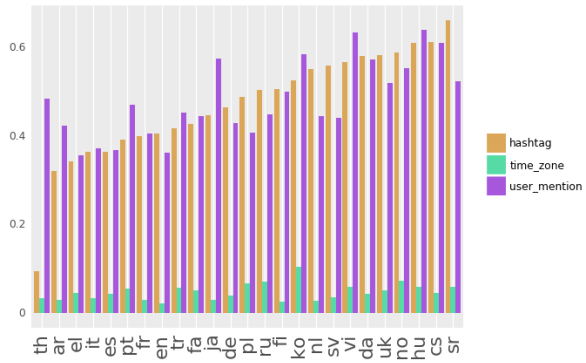

Figure 7: Jensen-Shannon Divergence calculated between GT and NGT hashtag, user mention, and time zone distributions, per language

The Spearman correlations between the variations are shown in Table 2. The values all indicate a positive association, but at a much lower level than the English role distributions. User mention and hash tag variations are more correlated with each other than either with time zone, which may be due to their intentional use compared to the passive setting of time zone by user devices (again, a better understanding of how time zones are set would help with interpreting this). [5] An interesting question for future work is whether the variations correlate with factors outside the scope of Twitter, such as government-driven propaganda, internet infrastructure, or cultural norms.

	hashtag	time_zone
user_mention	0.733	0.608
hashtag		0.638

Table 2: Pairwise Spearman correlation between JSD based on different distributions

5 Conclusion

We expanded previous work on differences between geo-tagged and non-geo-tagged English-language tweets to the full set of observed languages. In pursuit of aggregate user statistics, we determined that keyword search over user descriptions provides higher precision and recall than regular expressions applied to messages. We plan to exploit this further as supervised input to discrim-

[5] Note that the low divergences of the time zone distributions are likely because there is a strong correlation between the aggregate distributions of languages and time zones, while specific content (a political campaign, high-profile event, etc) can be very localized, and/or draw global interest.

inative models for extracting unconstrained self-identification in future work, and experiments on extending the method beyond English. Other interesting extensions include exploring correlations between the regional and language-specific variation and known cultural and political axes, and additional indexing of structure/content to compare other modes of variation. Finally, this study did not directly examine *content* fields (tweet texts and user descriptions) beyond the special case of role-extraction to generate additional categorical fields for English. Future work could extend it to variation over simple lexical features, which are easily extracted without language-specific processing.

References

Aragats Amirkhanyan and Cristoph Meinel. 2016. Analysis of the Value of Public Geotagged Data from Twitter from the Perspective of Providing Situational Awareness. *Social Media: The Good, the Bad, and the Ugly* .

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 *The Annual Meeting of the Association for Computational Linguistics (ACL)*. http://www.aclweb.org/anthology/P14-2030.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 1st edition.

Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H. Eugene Stanley. 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E* 65(4).

Diansheng Guo and Chao Chen. 2014. Detecting Non-personal and Spam Users on Geo-tagged Twitter Network. *Transactions in GIS* 18(3):370–384. https://doi.org/10.1111/tgis.12101.

M. Hollander, D.A. Wolfe, and E. Chicken. 2013. *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com/books?id=-V7jAQAAQBAJ.

Habib Karbasian, Hemant Purohit, Rajat Handa, Aqdas Malik, and Aditya Johri. 2018. Real-Time Inference of User Types to Assist with more Inclusive and Diverse Social Media Activism Campaigns. *Association for the Advancement of Artificial Intelligence* .

Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. 2015. Population Bias in Geotagged Tweets. *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop* .

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data. *CoRR* abs/1506.02275. http://arxiv.org/abs/1506.02275.

Luke Sloan. 2017. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society* .

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):1–20. https://doi.org/10.1371/journal.pone.0115545.

Stephen Wan Sunghwan Mac Kim and Cecile Paris. 2016. Detecting Social Roles in Twitter. *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media* .