The Effects of Data Collection Methods in Twitter

Sunghwan Mac Kim, Stephen Wan, Cécile Paris, Brian Jin and **Bella Robinson**
Data61, CSIRO, Sydney, Australia
{Mac.Kim, Stephen.Wan, Cecile.Paris, Brian.Jin, Bella.Robinson}@csiro.au

Abstract

There have been recent efforts to use social media to estimate demographic characteristics, such as age, gender or income, but there has been little work on investigating the effect of data acquisition methods on producing these estimates. In this paper, we compare four different Twitter data acquisition methods and explore their effects on the prediction of one particular demographic characteristic: occupation (or profession). We present a comparative analysis of the four data acquisition methods in the context of estimating occupation statistics for Australia. Our results show that the social network-based data collection method seems to perform the best. However, we note that each different data collection approach has its own benefits and limitations.

1 Introduction

Over the last decade, social media platforms have become prominent online channels for community interaction and communication. As a public data source, social media offers the potential to provide a cheap and large volume of real-time data to assist with social science research. Consequently, there have been several recent efforts to estimate aggregate demographic characteristics from social media (Sloan et al., 2015; Preoţiuc-Pietro et al., 2015) or to understand public views on topics like vaccination (Broniatowski et al., 2016). In such work, social media can supplement traditional data sources for social science research, such as interview and questionnaire data.

While different approaches to estimating demographic characteristics have been proposed, for example, for age and gender (Filippova, 2012) and for occupation (as these are useful as surrogates for income bracket) (Preoţiuc-Pietro et al., 2015), the effects of different data collection methods have been less studied. Twitter, as a source of predominantly public broadcast social media, allows for different methods for capturing user profile data, ranging from: (i) geolocation-based queries, (ii) word-based queries, (iii) Twitter's 1% sample stream, and (iv) social network-based crawling.

In this paper, we compare these four different Twitter data collection methods and explore their effects on estimating demographic characteristics. For this preliminary study, we focus on estimates of occupation groups for an Australian cohort and compare estimates to Australian 2011 census data.

We vary only the data collection method but use the same occupation statistic estimation throughout. We follow the methodology of Sloan et al. (2015), who use social media to estimate the United Kingdom (UK) occupation classes. This method requires an occupation taxonomy as the underlying resource for a keyword-spotting approach to compute the estimates. Sloan et al. (2015) used a resource called the *Standard Occupational Classification (SOC) 2010*[1], which was used to organise UK 2011 census data. As our estimates are for an Australian context, we use the corresponding *Australian*

[1] http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/index.html

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 86–91,
Austin, TX, November 5, 2016.

and New Zealand Standard Classification of Occupations Version 1.2 (2013) or *ANZSCO*, published by the Australian Bureau of Statistics (ABS).[2] The ABS used this resource to organise statistics from the 2011 census.

2 Data Collection Methods

This section describes the four data collection approaches employed to acquire sets of Australian Twitter user profiles: (i) Geo-location queries, (ii) word-based queries, (iii) Twitter's 1% sample, and (iv) social network-based crawling. The first three methods use data sourced from existing projects that collect Twitter posts. To remove time as a confounding factor, we used the largest intersection of collection periods, from April 1 to October 30, 2014.[3]

Geo-located data was sourced from the CSIRO Data61 Emergency Situation Awareness project (Cameron et al., 2012). In this project, we took Twitter data collected for the Australia and New Zealand region. The system, which focuses on event-detection for natural disasters, uses a series of latitude/longitude coordinates and a radius with Twitter's location-based Search API to define collection boundaries that cover the heavily populated regions in Australia and New Zealand. This system relies on Twitter's built-in functionality to infer location based on Twitter metadata. We refer to data collected via this method as the *Geo-location* method.

For **word-based queries**, the data collection was based on queries curated by the State Library of New South Wales (SLNSW) as described in (Barwick et al., 2014). The SLNSW has a mandate to collect and archive data about daily life in the Australian state of New South Wales (NSW). Since 2012, their collection has extended beyond traditional media (e.g., print newspapers) to include social media. Library staff curate a set of queries on a daily basis, reacting to the salient NSW-specific news of the day. This can thus span any news topic, including politics, government, arts, festivals, sports. To date, over 1000 queries have been curated in this fashion since 2012, including general hashtags for politics (e.g. "#auspol"), event specific queries (e.g. "Vivid Festival"), and personalities. We refer to data collected via this method as the *Word-based* method.

For the third method, we used the **1% Twitter sample** which was collected as part of the CSIRO Data61 WeFeel project (Larsen et al., 2015). This sample, colloquially known as the Spritzer stream, was used for studying the emotion content in Twitter to further research in mental health. We refer to data collected via this method as the *Spritzer* method.

The **social network-based crawling** method starts with a seed set of known Australian Twitter user profiles and crawls the social network multigraph of followers to find other Australian user profiles (Dennett et al., 2016). The seed set consisted of public celebrities, politicians, journalists, government accounts, and accounts for Australian companies and institutions. Each new user profile encountered during the crawling process was automatically labelled as being Australian using the location and timezone metadata together with a gazeteer of known Australian locations, and a label propagation method. For all discovered Australian accounts, the crawling process continued. A crawling depth of 3-hops was used from the seed accounts. We refer to data collected via this method as the *Social-network* method.

2.1 Data Pre-processing

For methods (i) to (iii), the corresponding user profiles for the authors of the collected tweets were also obtained using the Twitter API. All user profiles, regardless of method were filtered as follows. We first filtered accounts using an in-house text classifier on the profile user name and description to determine if the account represented an individual or an organisation, where it is the former that is of most interest for estimating demographic statistics. This classifier uses a maximum entropy model (Berger et al., 1996) for the binary distinction, *individual* versus *organisation*, which has an accuracy of 95.2%. Finally, Twitter metadata was used to further filter user profiles, keeping only those with an Australian time zone and English specified as the language.

2.2 Data Descriptive Statistics

Table 1 shows the number of Twitter user profiles with a breakdown by Australian states, identified using time zone information. In Australia, each state

[2] www.abs.gov.au/ANZSCO

[3] This end date was chosen as Twitter's location-based Search API was not fully functional after this date.

Region	Geo-location	Word-query	Spritzer	Social-network	Population
AU	624,769	66,812	202,657	873,899	$\approx 20 \times 10^6$
ACT	14,157	2,585	6,885	39,193	357,222
NSW	240,055	25,923	60,119	264,235	6,917,658
NT	6,530	356	1,450	6,509	211,945
QLD	119,858	14,028	52,514	217,744	4,332,739
SA	31,494	3,768	13,840	58,857	1,596,572
TAS	11,027	903	2,548	11,671	495,354
VIC	162,037	15,815	47,815	210,585	5,354,042
WA	39,611	3,434	17,486	65,105	2,239,170

Table 1: Number of Twitter user profiles for Australian individuals and census population for Australia and its states. Abbreviations: Australia (AU), Australian Capital Territory (ACT), New South Wales (NSW), Northern Territory (NT), Queensland (QLD), South Australia (SA), Tasmania (TAS), Victoria (VIC), Western Australia (WA).

has a different Twitter time zone setting based on the capital city for that state. The table also shows population statistics obtained from the 2011 census.

3 The ANZSCO Hierarchy

The ANZSCO hierarchy organises occupations into five levels of occupation categories. The top level, known as the *major group*, contains 8 occupation groups: managers, professionals, technicians and trades workers, community and personal service workers, clerical and administrative workers, sales workers, machinery operators and drivers, and labourers. Each major group is divided into sub-major groups, which are further divided into minor groups. Each minor group is divided into unit groups, which contain the leaf level specific occupations. The ANZSCO hierarchy has 8 major groups, 43 sub-major groups, 97 minor groups and 358 unit groups. There are 1,034 occupation names represented at the leaf level of the hierarchy. In this work, our correlations will be based on data for the major groups.

4 Estimating Occupation Statistics

Our aim here is to calculate the proportions for each occupation class at the major group level. We use the ANZSCO resource to provide a list of keywords to spot. These are derived from the node labels at each level in the hierarchy.

For any given collection of user profiles and the descriptions contained therein, when a match is found to a word in this list, a counter for the node responsible is incremented. We refer to this as our KeyWord Spotting (KWS) method[4], which is inspired from the methods described in (Sloan et al., 2015). As our evaluation uses the highest major group level, we propagate counts up through the hierarchy and sum them at the top level of the hierarchy. Finally, frequencies are normalised by the sum of frequencies over all 8 occupation categories to provide percentages, as in the census data. For the KWS method, words that occur under multiple categories at the major group level were discarded. For words that occurred in multiple nodes within a single branch of the major group, the highest level node was chosen to increment the counter. We performed text pre-processing prior to calculating the estimates in order to mitigate the noisiness of free text Twitter user profile descriptions. We removed non-ASCII characters and stop words, and all tokens were lower-cased. It is possible that multiple occupations are listed in a single user profile description. In this work, the first occupation word found is selected under the assumption that it is likely to represent the main occupation (Sloan et al., 2015).

Finally, we assembled the subsets of the Twitter user profiles, where occupations were identified using the KWS method. The number of profiles from each data collection method with a matching occupation is as follows: Geo-location: 100,829 / 624,769 (16.14%), Word-query: 16,358 / 66,812 (24.48%), Spritzer: 36,034 / 202,657 (17.78%) and Social-network: 104,867 / 873,899 (12.00%).

5 Comparisons to Census Data

In this evaluation, we look at the ranking of major group occupation categories based on social media estimates of prevalence and compare this derived ranking to the ordering from the 2011 census data. We used Kendall's τ (Kendall, 1938), a nonparametric statistical metric for comparing different rankings.

We calculate the Kendall τ rank correlation coefficient to compare the census occupation group percentages with the corresponding Twitter-derived

[4]While there has been a significant work on occupation inference (Preoţiuc-Pietro et al., 2015), we take a simple KWS approach to identify user occupations. Note that the primary goal of this work is to compare different data collection methods to estimate occupation statistics.

Region	Geo-location		Word-query		Spritzer		Social-network	
	cor	p-value	cor	p-value	cor	p-value	cor	p-value
AU	0.5714	0.0610	0.5714	0.0610	0.5714	0.0610	0.5714	0.0610
ACT	**0.7857**	0.0055	**0.7638**	0.0088	**0.7857**	0.0055	**0.7638**	0.0088
NSW	**0.7143**	0.0141	**0.7857**	0.0055	**0.7857**	0.0055	**0.7143**	0.0141
NT	0.5000	0.1087	**0.6183**	0.0341	0.5714	0.0610	**0.6429**	0.0312
QLD	0.5000	0.1087	0.4286	0.1789	0.4286	0.1789	0.4286	0.1789
SA	0.5000	0.1087	0.4728	0.1051	0.5000	0.1087	0.5714	0.0610
TAS	0.3571	0.2751	0.4286	0.1789	0.4001	0.1702	0.2857	0.3988
VIC	**0.6429**	0.0312	0.5000	0.1087	0.5714	0.0610	**0.6429**	0.0312
WA	0.5000	0.1087	0.4286	0.1789	0.4286	0.1789	0.4286	0.1789

Table 2: Kendall correlations for estimates of national and state occupation statistics derived by the KWS tagger. Bold indicates statistically significant results ($p<0.05$).

percentages from each data collection method. Table 2 shows the correlation coefficients for Australia and its states with respect to the Geo-location, Word-query, Spritzer and Social-network based methods. For determining significance, we set $\alpha = 0.05$.

We observe that the correlations are almost but not quite statistically significant at national level, with $p \approx 0.06$. We note that the correlation are identical for the national level. In this case, each method is resulting in the same number of ranking mistakes. As Kendall's τ is a measurement of the number of pairwise swaps needed to convert compare two rankings, the coefficients are identical

At the state-level, we observe that the Social-network data has the most states with significant correlations: 4 out of 7 states.[5] The Geo-location and Word-query based methods both have 3 states with significant correlations, whereas the Spritzer method has 2 states. This suggests that the social network crawling method performs better than the others at producing these estimates.

6 Discussion and Future work

Our results show that, for statistics at the national level, all methods appear to perform identically. However, for statistics at the state level, differences in the different data collection methods become apparent.

The Social-network method may be superior to the Spritzer method because it acquires a far larger set of user profiles. The same can be said about the Geo-location method which also collects a large number of Australian user profiles. This extra data, or the ability of the Social-network and Geo-location based methods to sample relevant profiles, results in significant correlations for VIC, the second most populous state in Australia, which is not captured well by the Spritzer method.

Interestingly, the Word-query method retrieves the smallest number of unique user profiles but does surprisingly well compared to the Spritzer method. We suspect this is due to the curation of queries that collect social media related to the state of NSW. Indeed, the correlation for NSW for this method is better than that of the Social-network approach. Furthermore, NSW has the highest correlation among all the states. We do note, however, that this method requires human-curated queries, a process that is time intensive.

For all methods, there are significant correlations for the ACT state. We find the ACT to be well represented in all of the social media collection methods, perhaps because it is the capital of Australia.[6] Presumably, a large volume of Twitter traffic is generated by government and industry staff located within the state. The Word-query method shows a significant correlation for the NT state. We suspect that the Word-query based method also does well for non-NSW states because the library uses some general queries like *#auspol*, which capture nation-wide discussions.

The Social-network method may have an advantage over the other data collection methods as it does not require users to actively post Twitter messages. Some Twitter users follow accounts of interest and rarely post messages themselves and therefore will be missed by the Geo-location, Word-query and Spritzer methods.

In this work, Kendall's τ coefficient does not provide deep insight at the national level of Australia. This is likely due to the number of categories being ranked. In the major group of the ANZSCO taxonomy, there are only 8 groupings of occupations. To provide further insights about the rankings at the national level, we visualise the major occupation rankings amongst the four data collection methods for Australia, as shown in Figure 1. The English letters on the X-axis correspond to 8 major occupations in

[5] Technically, the ACT and NT are territories, not states.

[6] Note that the ACT is geographically surrounded by the state of NSW.

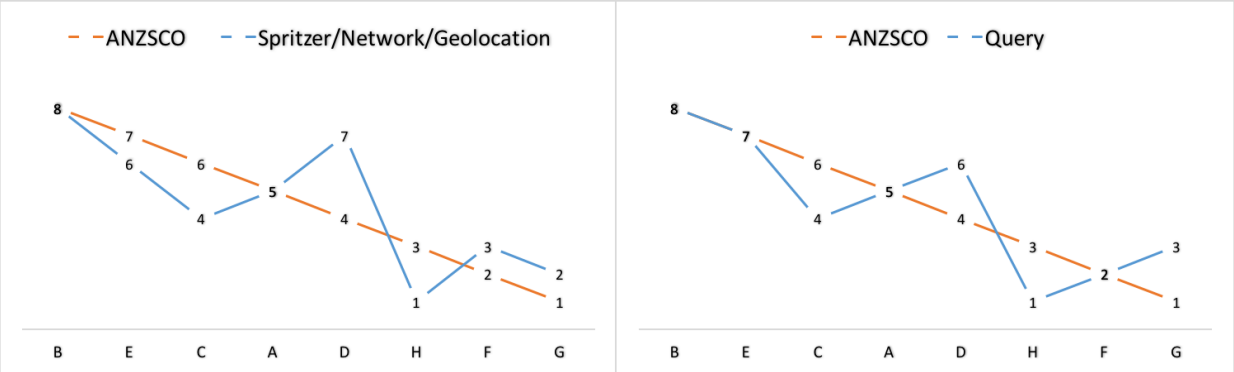

Figure 1: Comparison of major occupation rankings between ANZSCO and four data collection methods for Australia (A: Managers, B: Professionals, C: Technician and Trades Workers, D: Community and Personal Service Workers, E: Clerical and Administrative Workers, F: Sales Workers, G: Machinery Operators and Drivers, H: Labourers).

the ANZSCO hierarchy. These are listed according to ANZSCO rankings, with B at the highest rank and G at the lowest. The digits on the graph indicate the rankings produced by each data collection method. We notice that *Professionals* (B) and *Managers* (A), as the first and fourth ranked occupation groups in ANZSCO, are ranked correctly by all methods. Interestingly, the Word-query based method is the only one to correctly rank the *Clerical and Administrative Workers* (E) and *Sales Workers* (F) classes. We can only hypothesise that, because this method uses the queries capturing discussions about everyday life, it is able to better represent these subgroups.

The current study does have some limitations. One of these is that our Word-query method uses queries specific to one state in Australia, NSW, whereas the other data collection methods do not suffer from this bias. In future work, we will try to repeat our exploration of Word-query methods with a more general set of human-curated queries. We have also focused here on estimating statistics about occupation. We are also interested in examining the effects of data collection methods in estimating other demographic characteristics, such as age and gender. Finally, we would also like to replicate this work for other languages and countries outside of an Australian context.

7 Conclusion

In this paper, we see that different data collection methods have an effect on the quality of estimates of occupation classes. The question of which is best may depend on the application context requiring the estimate of occupation classes. If the aim is to produce an estimate for the current population, the Social-network approach may be best as it is able to find a large volume of user profiles, with little manual intervention. However, for many applications there may be a time-based element. For example, to study public discussion corresponding to a social event or information campaign taking place at a certain time, one may want to use posts colected using the Geo-location or Word-query based methods to better target the most relevant audience or community. Our study shows that methods based on posts can still yield good estimates.

Acknowledgments

The authors are grateful to anonymous NLPCSS@EMNLP reviewers.

References

Kathryn Barwick, Mylee Joseph, Cécile Paris, and Stephen Wan. 2014. Hunters and collectors: seeking social media content for cultural heritage collections. In *VALA 2014: Streaming With Possibilities*.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

David A Broniatowski, Mark Dredze, Karen M Hilyard, Maeghan Dessecker, Sandra Crouse Quinn, Amelia

Jamison, Michael J. Paul, and Michael C. Smith. 2016. Both mirror and complement: A comparison of social media data and survey data about flu vaccination. In *American Public Health Association*.

Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 695–698, New York, NY, USA. ACM.

Amanda Dennett, Surya Nepal, Cecile Paris, and Bella Robinson. 2016. Tweetripple: Understanding your twitter audience and the impact of your tweets. In *Proceedings of the 2nd IEEE International Conference on Collaboration and Internet Computing*, Pittsburgh, PA, USA, November. IEEE.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of EMNLP-CoNLL*, pages 1478–1488, Jeju Island, Korea, July. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Mark E. Larsen, Tjeerd W. Boonstra, Philip J. Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. 2015. We Feel: Mapping Emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1246–1252.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proceedings of ACL-IJCNLP*, pages 1754–1764, Beijing, China, July. Association for Computational Linguistics.

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, 03.