Using Twitter for public health infoveillance: a feasibility study

Andrew Jull[1], Adam Bermingham[2], Ayokunle Adeosun[2], Cliona Ni Mhurchu[1], Alan F. Smeaton[2]

[1]University of Auckland, Auckland, New Zealand
[2]Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland
Contact Alan.Smeaton@DCU.ie

Polls have been used for decades as a tool to gauge public opinion on specified topics. Originally used in US Presidential elections, they are now used to gather point-of-time information on politicians, political issues, brand names, products, even prospective storylines in movies. Robust polls typically specify *a priori* an acceptable level of sampling error e.g. +/- 3% in order to calculate the size of the random sample needed. While these error rates are acceptable for most cases, having to manually poll a random sample of up to 1,000 people means opinion polls are expensive to carry out. In order to identify secular trends, polling must be repeated at frequent intervals, which makes monitoring for trends even more expensive. Surveys can prove more cost-effective, but still require resources to recruit participants and collate results, with the risk of latency from time of issuing a survey to aggregation of results. For this reason we look to online social media as a potential low-cost, continuous and scalable alternative source for opinion mining that can be readily replicated.

There are very few studies analyzing tweets to analyse the demography of those tweeting or studies which have considered how representative sampled tweets may be, something integral to traditional opinion polling. Sampling bias too may influence Twitter analyses as there is anecdotal evidence of "gaming" tweets by sentiment in, for example, political elections. Thus we focus on analysis from a sub-population of those who tweet and the degree to which such a cohort of Twitter accounts could be representative of the population of users.

Using Twitter as a source of opinion has typically treated all tweets as the population, sampled over periods of time which may be useful for example for influenza infoveillance as in Google Flu Trends, prior to its recent demise. Our premise is that if a tweet mentions a topic, it bears sentiment (+, - or neutral) with reference to that topic. Since the volume of data on social media is very large, we can use statistical techniques such as machine learning to automate the determination of sentiment. Sentiment analysis is just one of many data-driven statistical techniques which can be used to aggregate unstructured textual data, but training these systems for a given domain is contingent on the availability of large volumes of relevant data. Language and vocabulary differences make repurposing out-of-domain models problematic.

Developments in Ireland and in New Zealand offer test-case opportunities for public health infoveillance. Smoking and diet-related risk factors are known leading risk factor contributors to the global burden of disease.[1] Both Ireland and New Zealand have taken bold steps to reduce smoking prevalence, but have chosen different methods for tackling obesity despite rates of obesity and overweight (61% Irish adults and 22% Irish children are overweight/obese compared with 65% of New Zealand adults and 33% of children).[2 3] New Zealand has been focused on individual responsibility rather than government leadership and intervention in contrast to Ireland which supports development of personal skills and addressing environments to support healthier choices.

Within this domain we set out to evaluate two inter-related concepts within infodemiology, a developing field that brings together epidemiology and information technology.[4] First we sought to investigate the feasibility of accumulating a cohort of Twitter accounts specific to the two countries, and to characterize and study the sample cohorts. Second, we sought to investigate the feasibility of accumulating and analyzing tweets related to four defined areas of public health interest, namely: *fast foods, sugar-sweetened beverages, energy drinks* and *e-cigarettes* and explore the related sentiments in both the study and the sample cohorts.

Methods

Derivation of the study, DataSift and random sample

We used an online data aggregation platform (www.datasift.com) to identify and crawl our study cohort accounts. We set the collection period to be long enough that active Twitter accounts were captured, and not so short that the set is influenced by any particular news event in either country. We constructed a location-based historic query of tweets from New Zealand and from the Republic of Ireland that collected the 200 most recent tweets from these accounts for the two months from 16th June to 16th August 2014. It was not possible to crawl the entire set of the study cohort as Twitter API requests may be rejected for reasons that include a user having changed their account to private, been suspended from Twitter, or a variety of unpredictable issues related to Twitter API issues.

In order to derive a study cohort of active human Twitter users, we set filtering criteria based on tweet frequency in our initial sample to remove dormant accounts and hyperactive accounts. We defined an active human account to be one where the user tweeted on average at least once every 14 days and a hyperactive account to be one where the user tweeted more than 100 times per day. We assumed a hyperactive account to be a non-human user or group account. We then randomly sampled the study cohort to derive a sample cohort.

Derivation of location and profile characteristics

We used two criteria to identify location. First we sought tweets where the author's profile location contained the country name i.e. "Ireland" or "New Zealand". The second criterion was a specified geo-coordinate bounding box around Ireland and New Zealand. If a tweet was found to match either of these criteria, we added that tweet to the candidate tweet pool, for that country. After our initial data collection we further refined the dataset by removing tweets from Northern Ireland. To achieve this refinement, we filtered our data using the Twitter "tweet/place/country_code" field attribute that Twitter provides as a reverse geocoding of the tweet coordinates.

DataSift provides classification of profile gender based on comparison of the profile name to a set base of names with known gender. We aggregated "mostly_female" and "female" to "female", and similarly "mostly_male" and "male" to "male". We described as "human" the union of the aforementioned conflated male and female sets, and with the relatively infrequent "unisex" classification. This approach is a precision-oriented method of determining gender whereby profiles are unlikely to be confused between the genders and non-human profiles are unlikely to be included in the male, female or human sets.

Box 1. Domain names with related topic names

Sugar-sweetened soda: Coke, coca cola, CocaColaCo, WorldofCocaCola, CocaCola, ACocaColaCo, DietCoke, Cokejourney, CokeZero, CokeZone, CherryCoke, VanillaCoke, ShareACoke, CokeZeroBikes, Dr Pepper, Dr. Pepper, DrPepper, Sprite, SpriteZero, Diet Sprite, Foxton Fizz, Lemon & Paeroa, 7up, Pepsi, Diet Pepsi, Pepsi Max, Dietpepsi, PepsiMaxUK, PepsiMAX, PepsiCola, PepsiCo.

Energy drinks: Red Bull, redbull, redbullUK, VenergydrinkAUS, VEnergyNZ, LucozadeEnergy, Lucozade, Monster Energy, MonsterEnergy, redbullire, MonsterEnergyIE, MonsterEnergyAU

Fast food: Burger King, BurgerKing, kfc, Love_KFC, kentucky fried chicken, Abrakebabra, Supermac's, SupermacsIRE, Nando's, NandosUK, NandosIRE, Nandos, quiznos, quizno's, mcdonalds, mcdlimerick, mcdonald's, maccasnz, mcdonalds, mcdonaldseurope, dominos pizza, dominos_roi, pizza hut, pizzahut, pizzahutdeliver, pizzahutirl, pizzahutnz, subwayukireland, eddie rockets , eddierocketsirl , subwaynz, wendy's, pitapit, pita pit, pitapitnz

E-cigarettes: v2cigs, V2 Cigs, VaporZoneMIA, VaporZone, HaloCigs, Halo Cigs, BullSmoke, Bull Smoke, Apollo E-Cigs, apolloecigs, blucigs, Blu Cigs, EverSmoke, Ever Smoke E-Cigs, JoyetechUK, Joyetech E-Cigs, SmokeFreeDN, Smoke Free E-Cigs, EonSmoke, electronic cigarette, e-cigarette, vaporizer, vaporiser, electronic vaping device, ecig, EUecigBan, NO2EUECIGBAN, vapeworld, ecigaretteforum, vape, vaping, vapelife, halocigs, VaporZone, eliquid.

Statistical analysis

Simple descriptive statistics were generated to describe demographic characteristics for each cohort. Medians and interquartile ranges (IQR) were reported where data was non-parametric. We defined four domains (Box 1), namely sugar sweetened soda, energy drinks, fast food and e-cigarettes, and within each domain, we defined topics using key word descriptors to be contained in tweets, (branded names and where appropriate shortened brand names e.g. Coke), corporate account names and hashtags relevant to Ireland and New Zealand. Sentiment analysis was then undertaken.

Results

Although a similar proportion of the study cohorts were considered human accounts (Tables 1 and 2), more accounts in the Irish study cohort than in the New Zealand study cohort could be crawled during the study period to derive the DataSift cohorts for each country (57.8% versus 26.7% respectively). These proportions shifted somewhat when only human accounts were the denominator, but the Irish cohort still represented a larger percentage of the human accounts than the New Zealand cohort (78.6% versus 38.8% respectively). Thus the random samples of 5,000 human accounts represented about one in three of the DataSift human accounts in the New Zealand sample, but only about one in 11 DataSift accounts in the Irish sample. Despite these differentials, a sample of 5,000 Twitter accounts was similarly representative of the DataSift cohort in both the Irish and New Zealand populations. The largest difference between the random sample and the DataSift cohort on any one demographic characteristic was on Accounts with a location profile in both the New Zealand and Ireland samples (3.4% and 3.7% differences respectively).

The patterns of tweets were the same in both the New Zealand and Irish DataSift cohorts, both with respect to the number of accounts tweeting about the domains and the number of tweets about the domains. Fast food was most frequently tweeted, followed by sugar-sweetened beverages, and energy drinks. E-cigarettes were an infrequently tweeted domain. Fast food was tweeted by similar proportions of the populations in both the New Zealand and Irish DataSift cohorts (about 14%). In both the New Zealand and the Irish populations the random sample of accounts was representative with respect to the number of accounts with tweets about sugar-sweetened soda, energy drinks, fast food and e-cigarettes. The differences between the random sample and the DataSift cohort in the number of accounts tweeting about the domains were less than 1%. While the random samples seem to reflect the number of accounts tweeting about the domains, the volumes of tweets were under-represented, particularly in the Irish sample where the number of tweets in each domain were about half that of the DataSift cohort in each domain.

The sentiment analysis of the tweets from the two samples of 5,000 users shows that no matter what the domain, the great majority of terms are not associated with either positive or negative sentiment (Figure 1).

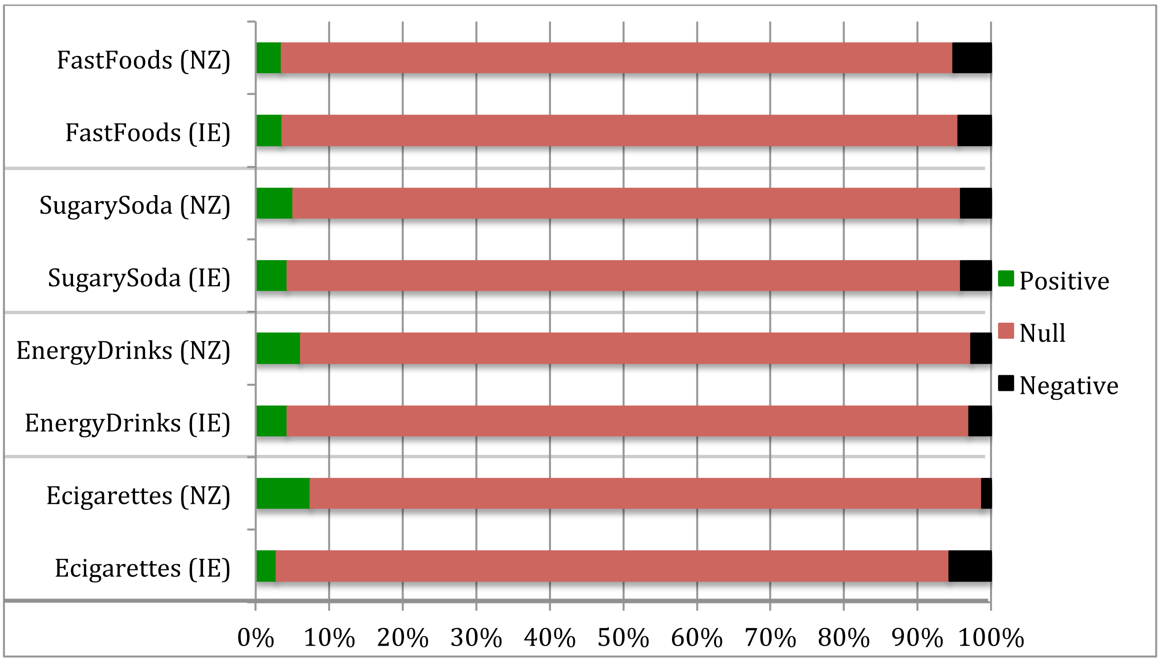

Figure 1. Sentiment analysis of terms used in tweets about fast foods, sugar-sweetened soda, energy drinks, and e-cigarettes from a random sample of 5000 New Zealand and Irish Twitter accounts.

Discussion

Principal Results

We have demonstrated that it is possible to assemble a cohort of Twitter accounts associated with specific countries using Twitter profiles, that we can accumulate tweets from a cohort of such accounts, and that randomly sampling 5,000 of these accounts creates a cohort representative of the larger cohort, both with respect to the account and the tweet characteristics available to us. It appears therefore that it is both possible and feasible to assemble a cohort of Twitter accounts over time in order to assess domains of interest associated with, in our case, non-communicable disease. Such follow-up could offer real-time cost-effective data on Twitter users' perceptions and the ability to undertake trans-national comparisons with ease in order to measure the effects of natural experiments, policy changes and new marketing practices on health-related behaviours.

Sentiment analysis has had limited application in healthcare. A simple OVID Medline search reveals 54 citations (accessed 1 March 2016) related to sentiment analysis or opinion mining and Twitter, with increased frequency of publication 2012-2015 (3, 9, 10, 26, and 4 publications in each year from 2012). While increasing attention has been given to public health topics (e.g opinions about vaccination and cancer screening programmes), investigations related to leading risk factors for non-communicable disease have largely been limited to tobacco control (four publications since 2013). To the best of our knowledge all of the investigations have used a sample of tweets, rather than sampled accounts and only one study has previously explored sentiment in the area of obesity in social media, although not addressing the risk factors associated with weight gain.[5] That study also used a commercial service to crawl for data and found that Twitter provided 92% of its social media data (~1.2 million tweets). Perhaps more importantly, Twitter differed from the other social media channels in being seen as a unique channel for perpetuating social belief, a finding that is supported by investigations into Twitter and alcohol consumption.[6] These findings are of particular import given the changing patterns of media use, with the potential for users to be more active participants in organic (i.e. non-commercial) promotion of risk behaviours via Twitter, as well as understanding how such information spreads. Following a cohort of accounts over time may identify influential nodes within networks of social accounts as people surround themselves in social media with those they like and to whom they are alike.

The frequency of accounts tweeting about fast foods, sugar-sweetened beverages and energy drinks suggests sampling accounts may be a satisfactory approach for these domains, the infrequency of tweets about e-cigarettes across both the DataSift cohort and the random sample suggests that sampling accounts for some topics may not offer the same surveillance possibilities as monitoring the Twitter firehose or a decahose. Using geo-located tweets representing about 1% of all tweets, approximately 20,000 tweets per month relating to tobacco and e-cigarette use were obtained for North America.[7] The volume of tweets was moderately correlated to positive sentiment, but with state variation, suggesting that sentiment could assist in identifying real-time market segments for public health messaging.

Limitations

This study is subject to one major limitation, namely that Twitter users are unlikely to be representative of the general population.[8] However, it was not our intent to use Twitter to assemble cohorts that represented the general population. It does appear that sampling 5,000 accounts creates cohorts that are representative of Twitter users, albeit on a limited set of characteristics. More work is needed to establish in detail how to best characterize Twitter users and establish how representative a sample may be, as well as how frequently to refresh a cohort as accounts become inactive. While these limitations are real, from a health perspective the value of these data is in their real-time nature, cost-effectiveness, and ability to undertake cross-national comparisons with ease (e.g. to measure effects of 'natural experiments, policy changes, or new marketing practices on health-related behaviours).

References

1. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380(9859):2224-60.
2. Ministry of Health. *New Zealand Health Survey: Annual update of key findings 2012/13.* Wellington: Ministry of Health, 2013.
3. Oireachtas Library & Research Office. *Obesity - a growing problem? Secondary Obesity - a growing problem?* 2011. http://www.oireachtas.ie/parliament/media/housesoftheoireachtas/libraryresearch/spotlights/spotObesity071111_150658.pdf
4. Eysenbach G. Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health. *Am J Prev Med* 2011;40(5S2):S154-S58.
5. Chou W-YS, Prestin A, Kunath S. Obesity in social media: a mixed methods analysis. *Translat Behav Med* 2014;4(3):314-23. doi: 10.1007/s13142-014-0256-1.
6. Cavazos-Rehg PA, Krauss MJ, Sowles SJ, Bierut LJ. “Hey Everyone, I’m Drunk.” An Evaluation of Drinking-Related Twitter Chatter. *J Stud Alcohol Drugs* 2015;76:635-643.
7. Clark E, M., Jones C, Gaalema D, et al. Social media meets population health: a sentiment and demographic analysis of tobacco and e-cigarette use across the "Twittersphere". *Value Health* 2014;17:A603.
8. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN. Understanding the demographics of Twitter users. *Proceedings of the Fifth Interntional AAAI Conference on Weblogs and Social Media*, 554-557.

Table 1. Characteristics of New Zealand Twitter accounts and tweets

	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	60784		
Human	41887 (68.9)	16259 (38.8)	5000 (11.9)
Male	18179 (43.4)	7304 (44.9)	2275 (45.5)
Female	18799 (44.9)	6902 (42.5)	2096 (41.9)
Unisex	4912 (11.7)	2053 (12.6)	629 (12.6)
Accounts with geocoded tweets (NZ)	9039 (14.9)	3584 (22.0)	1064 (21.3)
Accounts with location profile (NZ)	55098 (90.7)	14113 (86.8)	4508 (90.2)
Klout (Median, IQR)	21 (14 – 32)	25 (18 – 36)	27 (19 – 40)
Followers (Median, IQR)	67 (14 – 246)	157 (63 – 414)	182 (71 – 477)
Account tweeted about			
Sugar-sweetened soda	-	1146 (7.0)	305 (6.1)
Energy drinks	-	580 (3.6)	184 (3.7)
Fast food	-	2277 (14.0)	673 (13.5)
E-cigarettes	-	34 (0.20)	12 (0.24)
Tweets	1853984	707510 (38.2)	295924 (16.0)
Geocoded tweets	141529 (7.6)	58323 (8.2)	22732 (7.7)
Geocoded tweets (NZ)	129854 (7.0)	53719 (7.6)	21141 (7.1)
Tweets with account location profile	1834397 (98.9)	701525 (99.2)	293836 (99.3)
Tweets with account location profile (NZ)	1772722 (95.6)	674139 (95.3)	282931 (95.6)
Tweets about			
Sugar-sweetened soda	-	1391 (0.20)	375 (0.13)
Energy drinks	-	885 (0.13)	282 (0.10)
Fast food	-	3156 (0.41)	928 (0.31)
E-cigarettes	-	52 (0.01)	20 (0.01)

Table 2. Characteristics of Republic of Ireland Twitter accounts and tweets

	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	93400		
Human	68780 (73.6)	54027 (57.8)	5000 (5.4)
Male	30616 (44.5)	24367 (45.1)	2165 (43.3)
Female	31273 (45.4)	23763 (44.0)	2247 (44.9)
Unisex	6891 (10.0)	5897 (10.9)	588 (11.8)
Users with geocoded tweets (IE)	17859 (19.1)	11729 (21.7)	1006 (20.1)
Users with location profile (IE)	81686 (87.5)	46483 (86.0)	4484 (89.7)
Klout (Median, IQR)	28 (20 – 38)	30 (22 – 39)	33 (25 - 40)
Followers (Median, IQR)	227 (91 – 508)	277 (135 – 594)	318 (160 – 659)
Account tweeted about			
Sugar-sweetened soda	-	5296 (9.8)	518 (10.4)
Energy drinks	-	1971 (3.6)	170 (3.4)
Fast food	-	7853 (14.5)	698 (14.0)
E-cigarettes	-	160 (0.30)	11 (0.22)
Tweets	1016435	604846 (59.5)	105449 (10.4)
Geocoded tweets	131690 (13.0)	119835 (19.8)	13959 (13.2)
Geocoded tweets (IE)	118832 (11.7)	83974 (13.9)	12857 (12.2)
Tweets with user location profile (IE)	945025 (93.0)	559053 (92.4)	98170 (93.1)
Tweets about			
Sugar-sweetened soda	-	6378 (1.1)	623 (0.59)
Energy drinks	-	2988 (0.49)	220 (0.21)
Fast food	-	10696 (1.8)	991 (0.94)
E-cigarettes	-	208 (0.03)	11 (0.01)