Centre for
Data Analytics

Insight

Using Twitter for Public Health Infoveillance: A Feasibility Study

Andrew Jull(*), Adam Bermingham(+), Ayokunle Adeosun(+), Cliona Ní Mhurchu(*), Alan F. Smeaton(+)

(*) University of Auckland, NZ
(+)Dublin City University, IRL

So who are all these people ?

- Andrew Jull … Public health, nursing science
- Adam Bermingham … Technology, social media
- Ayokunle Adeosun … Summer intern
- Cliona Ní Mhurchu … public health nutrition, population dietary interventions
- Alan F. Smeaton … all things data !

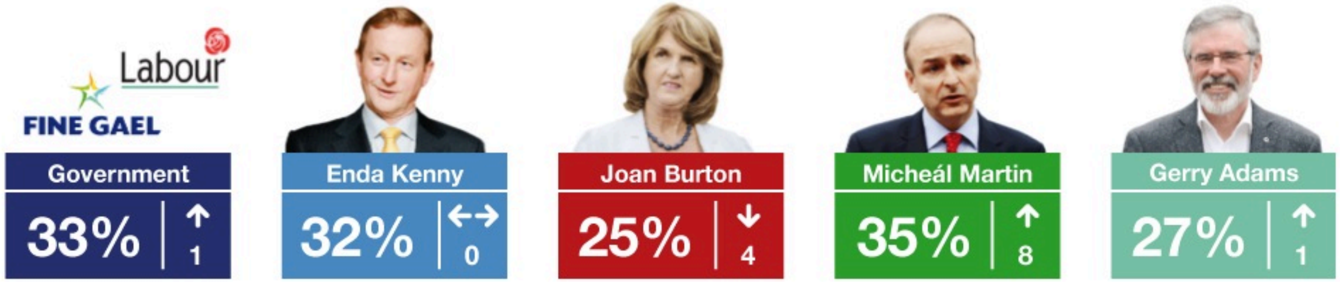
What's it about …

Twitter … Public Health … Infoveillance … Feasibility Study

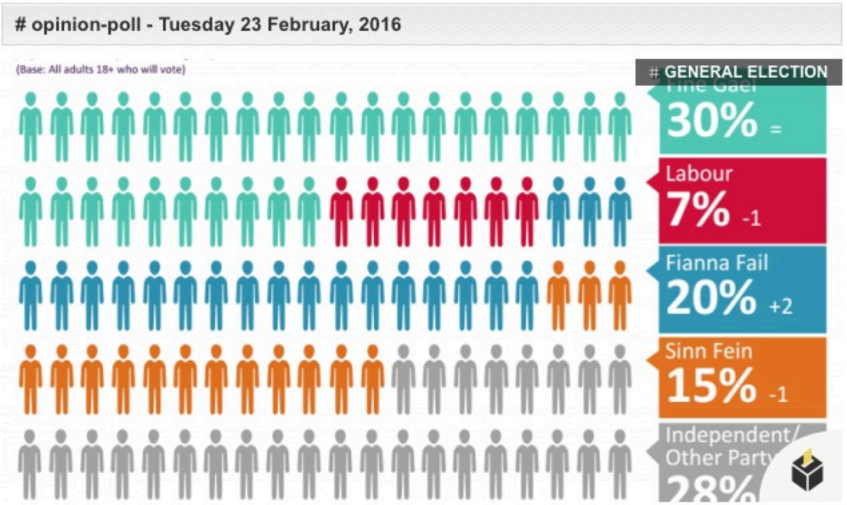
Polling

- Polls are informal tools used to gauge public opinion on topics
- Originally used in US Presidential elections, now used for point0in-time information on … politicians, political issues, brand names, products, movie storylines
- Robust polls have an acceptable sampling error, e.g. +/- 3%, in order to calculate the sample size needed
- Manually polling a truly random (gender, demographic, location, etc.) population of c.1,000 people means polls are expensive to carry out
- To identify secular trends, polls must be repeated at frequent intervals
- So, when you look at …

THE IRISH TIMES POLL | Ipsos MRBI POLL



thejournal.ie
READ, SHARE AND SHAPE THE NEWS


opinion-poll - Tuesday 23 February, 2016

Surveys

- Surveys are like polls but ask more questions thus are more in-depth
- Also require recruiting participants, address sample bias, collate results and suffer from latency between survey, and aggregation of results

- We look to online social media – Twitter – as a potentially …
  - Low cost
  - Continuous
  - Scalable
- … form of opinion mining that can be replicated

Previous Work

- Model political sentiment by mining social media … capture the voting intentions of a nation during an election campaign ?
- 2011 Irish General Election as a case study … sentiment analysis using supervised learning and volume-based measures.
- Evaluate against conventional election polls and final election result.
- Found social analytics (volume-based and sentiment analysis) are predictive .
- Observations related to monitoring public sentiment during an election campaign, including examining sample sizes, time periods as well as methods for qualitatively exploring the underlying content.

Bermingham, Adam and Smeaton, Alan F. (2011) *On using Twitter to monitor political sentiment and predict election results.* In: Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), 13th November 2011, Chiang Mai, Thailand.

Previous Work

- An unwritten finding was a question about how representative the tweets are !
- Few studies analyse the representative nature of tweets, apart from the fact there is a bias towards those who tweet anyway
- Post Obama Re-election 2012 we see how the candidate(s) easily use the medium of social media to promote messages, but is there “gaming” of followers, likes, re-tweets, etc., by bots ?
- Why not ? If there are bots (automated scripts that produce content and mimic real users) that play World of Warcraft then there could be bots that game “public” political sentiment ?
- A 2015 study of elections in Venezuela found governments and political actors make use of social bots, that fake social media accounts spread pro-governmental and anti-governmental messages, beef up web site follower numbers, and cause artificial trends.
- They believe that bot-generated propaganda and misdirection is a worldwide political strategy.
- Robotic lobbying tactics have been deployed in Russia, Mexico, China, Australia, the United Kingdom, the United States, Azerbaijan, Iran, Bahrain, South Korea, Turkey, Saudi Arabia, and Morocco.
-

Forelle, Michelle C and Howard, Philip N. and Monroy-Hernandez, Andres and Savage, Saiph, Political Bots and the Manipulation of Public Opinion in Venezuela (July 25, 2015). Available at SSRN: http://ssrn.com/abstract=2635800 or http://dx.doi.org/10.2139/ssrn.2635800

Related Work

- Sometimes the errorsome results in aggregation are not deliberate …
- Google Flu Trends since 2008 identified flu outbreaks by tracking users' searches about the symptoms and relief options … much earlier than the CDC
- However, with all the media coverage it focused people's attention on it, so results became skewed … especially when Google introduced auto-complete in search input

- Given that statistical techniques and machine learning can be used to determine tweet sentiment on some topic, and that this is used in everything from political elections to tracking brand marketing campaigns, we wonder how reliable is the underlying data, even if you take ALL the data ?
- So with these recent caveats, we're focusing not just on **WHAT** is tweeted, how often, and from where, but on **WHO** is tweeting.

Public Health Infoveillance surveillance using online information

- Smoking and diet are known global risk factors to the global burden of disease
- Ireland and New Zealand have taken similar steps to reduce smoking prevalence (ban, cost, warnings, under 18s, etc.) but different approaches to tackle obesity (61% Irish and 65% NZ adults are overweight, 22%/33% children)
- NZ government emphasised personal responsibility, Ireland emphasises environments to support healthier choices
- We set out to investigate 2 questions
    1. Can we accumulate an unbiased cohort of Twitter accounts for NZ and IRL?
    2. Can we accumulate tweets from these cohorts in 4 areas of public health interest?

- We used DataSift for a location-based historic query of tweets from NZ and IRL, gathering 200 most recent tweets from these accounts for 2 months during Summer 2014.
- Location based on author profile address and/or GPS-tagged tweets, and for NZ is even easier because of time zone.
- We filtered the users based on tweet frequency to remove dormant and hyperactive (bot ?) accounts, then randomly sampled the cohort to derive a sample of (real) users
- DataSift classifies user gender based on profile name vs a list of names
- We defined 4 health domains of interest, and for each we defined keyword topics, words to be contained within tweets

Terms for Four Health Domains

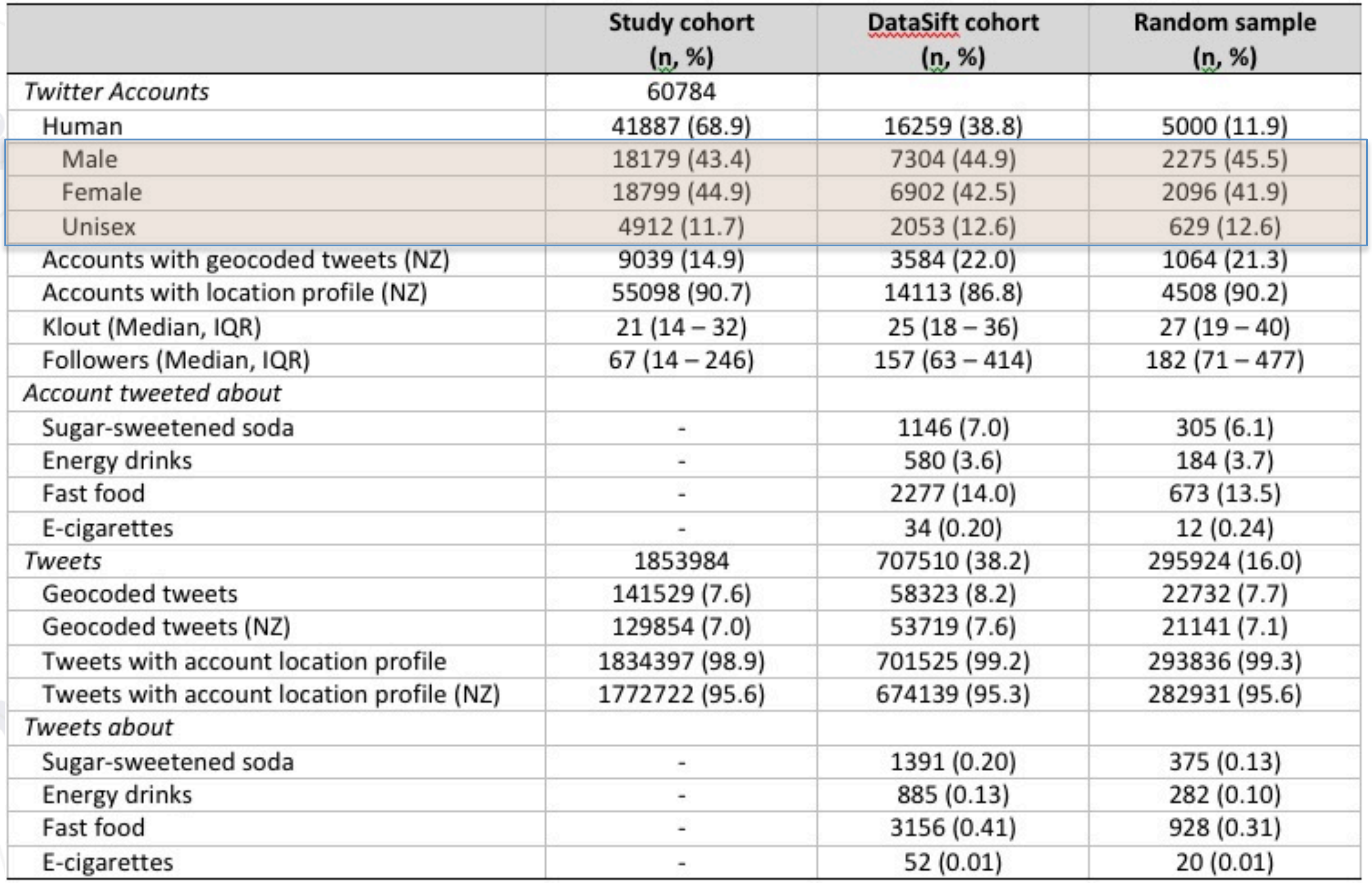
Sugar-sweetened soda: Coke, coca cola, CocaColaCo, WorldofCocaCola, CocaCola, ACocaColaCo, DietCoke, Cokejourney, CokeZero, CokeZone, CherryCoke, VanillaCoke, ShareACoke, CokeZeroBikes, Dr Pepper, Dr. Pepper, DrPepper, Sprite, SpriteZero, Diet Sprite, Foxton Fizz, Lemon & Paeroa, 7up, Pepsi, Diet Pepsi,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

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

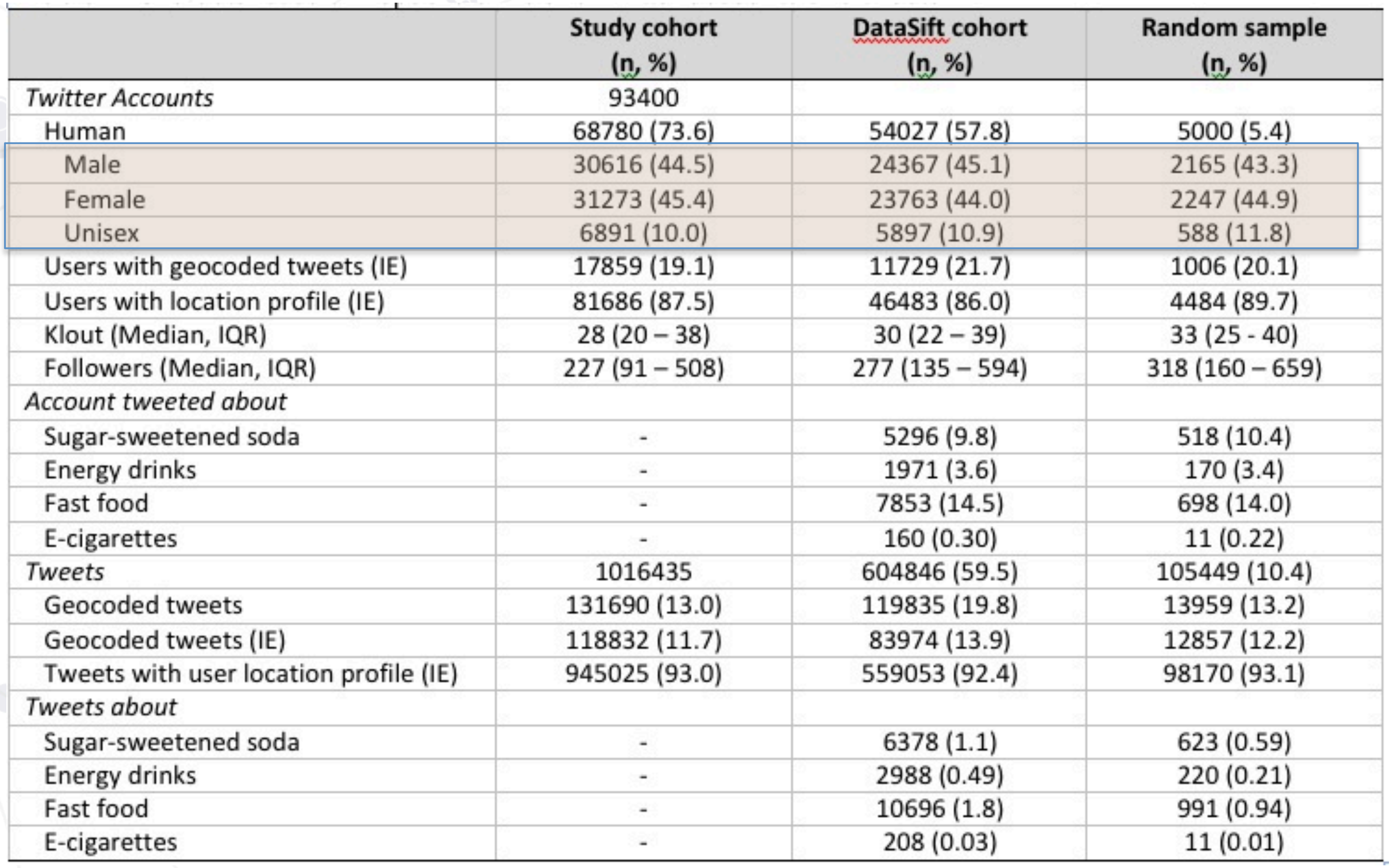
Fast food: Burger King, BurgerKing, kfc, Love_KFC, kentucky fried chicken, Abrakebabra, Supermac's, SupermacsIRE, Nando's, NandosUK, NandosIRE, Nandos, quiznos, quizno's, mcdonalds, mcdlimerick, mcdonald's, maccasnz, mcdonalds, mcdonaldseurope, dominos pizza, dominos_roi, pizza hut, pizzahut, pizzahutdeliver, 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.

NZ

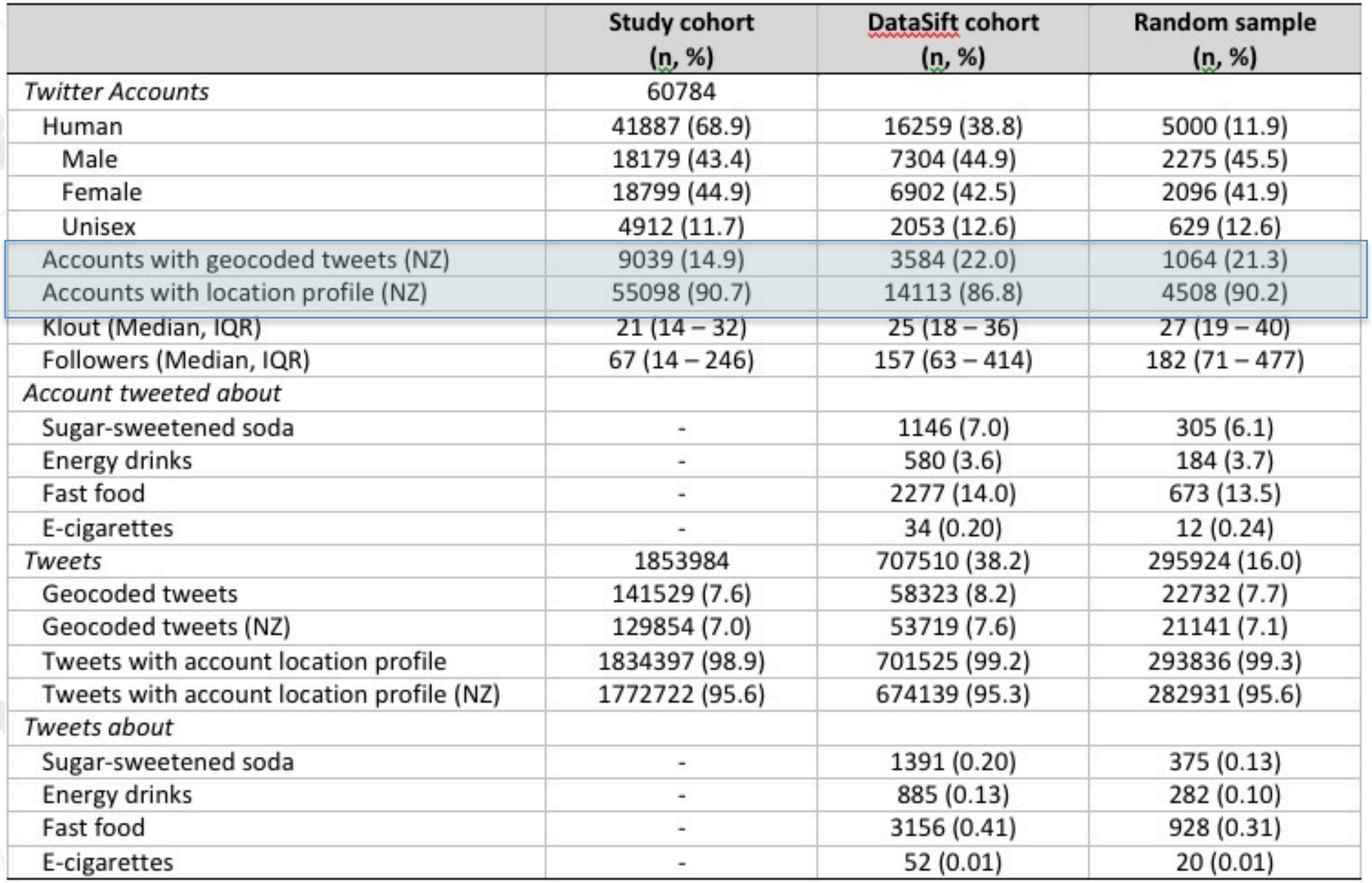


	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	60784		
Human	41887 (68.9)	16259 (38.8)	5000 (11.9)
Male	18179 (43.4)	7304 (44.9)	2275 (45.5)
Female	18799 (44.9)	6902 (42.5)	2096 (41.9)
Unisex	4912 (11.7)	2053 (12.6)	629 (12.6)
Accounts with geocoded tweets (NZ)	9039 (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)

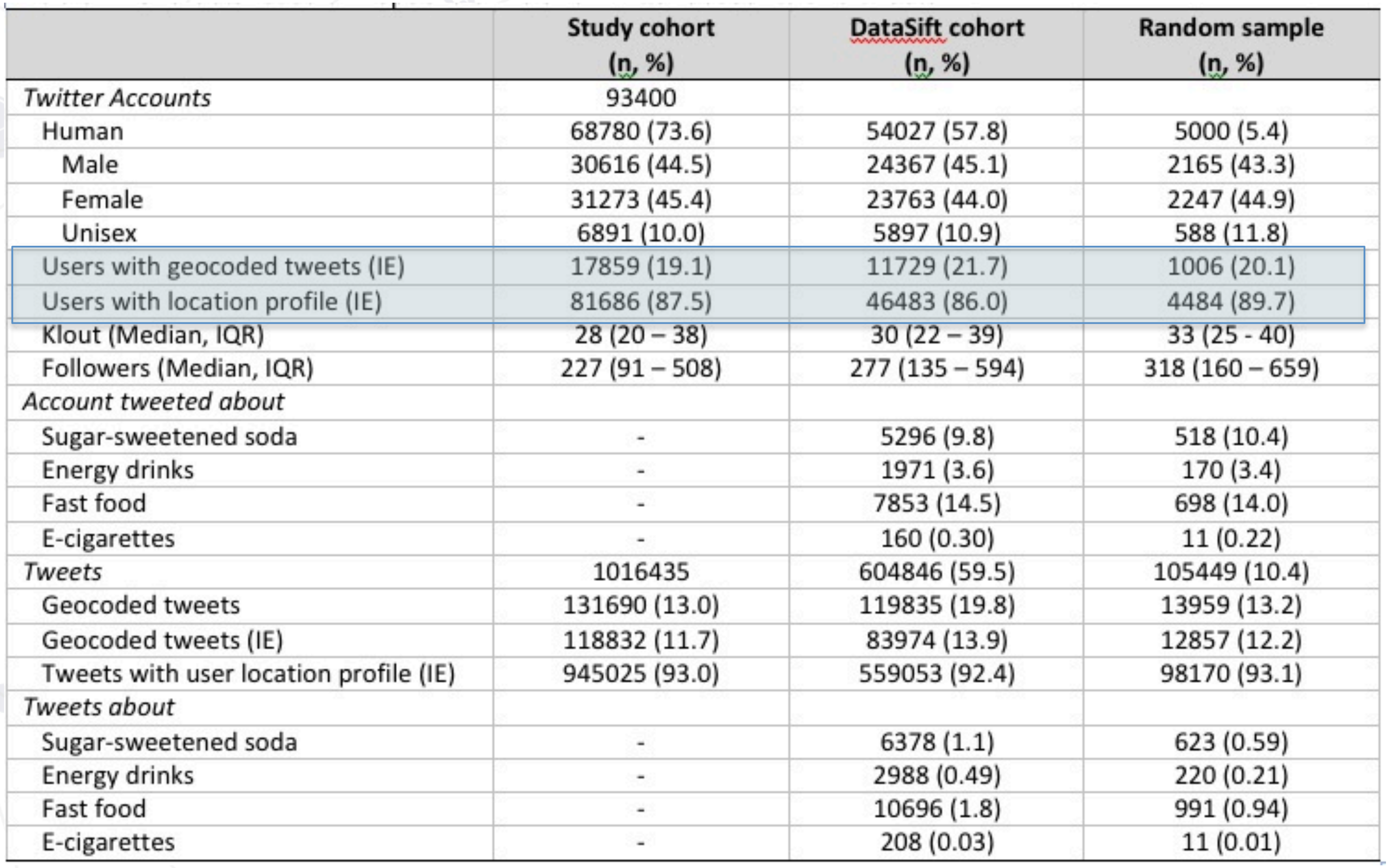


	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	93400		
Human	68780 (73.6)	54027 (57.8)	5000 (5.4)
Male	30616 (44.5)	24367 (45.1)	2165 (43.3)
Female	31273 (45.4)	23763 (44.0)	2247 (44.9)
Unisex	6891 (10.0)	5897 (10.9)	588 (11.8)
Users with geocoded tweets (IE)	17859 (19.1)	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)

NZ



	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)

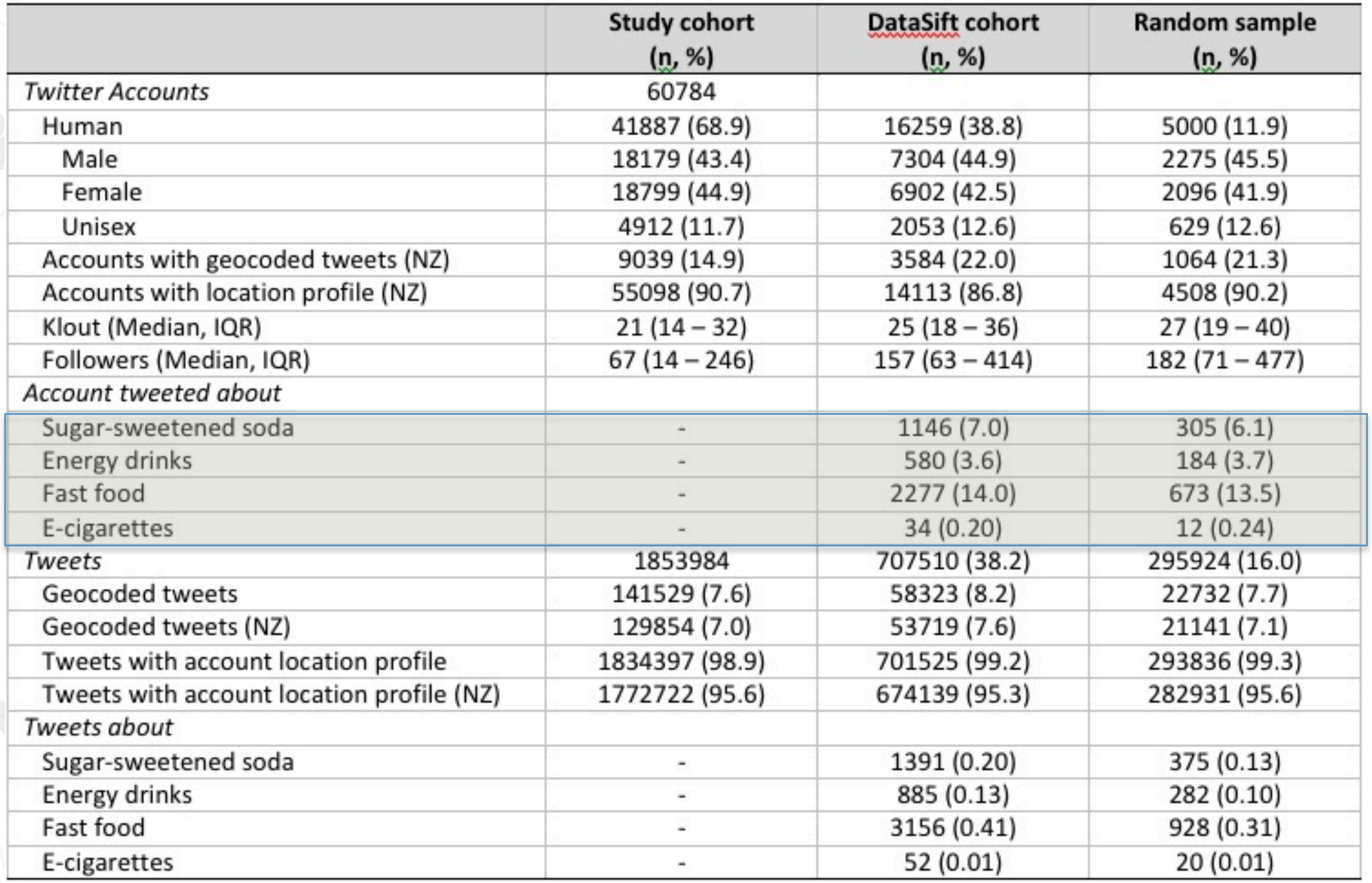


	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	93400		
Human	68780 (73.6)	54027 (57.8)	5000 (5.4)
Male	30616 (44.5)	24367 (45.1)	2165 (43.3)
Female	31273 (45.4)	23763 (44.0)	2247 (44.9)
Unisex	6891 (10.0)	5897 (10.9)	588 (11.8)
Users with geocoded tweets (IE)	17859 (19.1)	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)

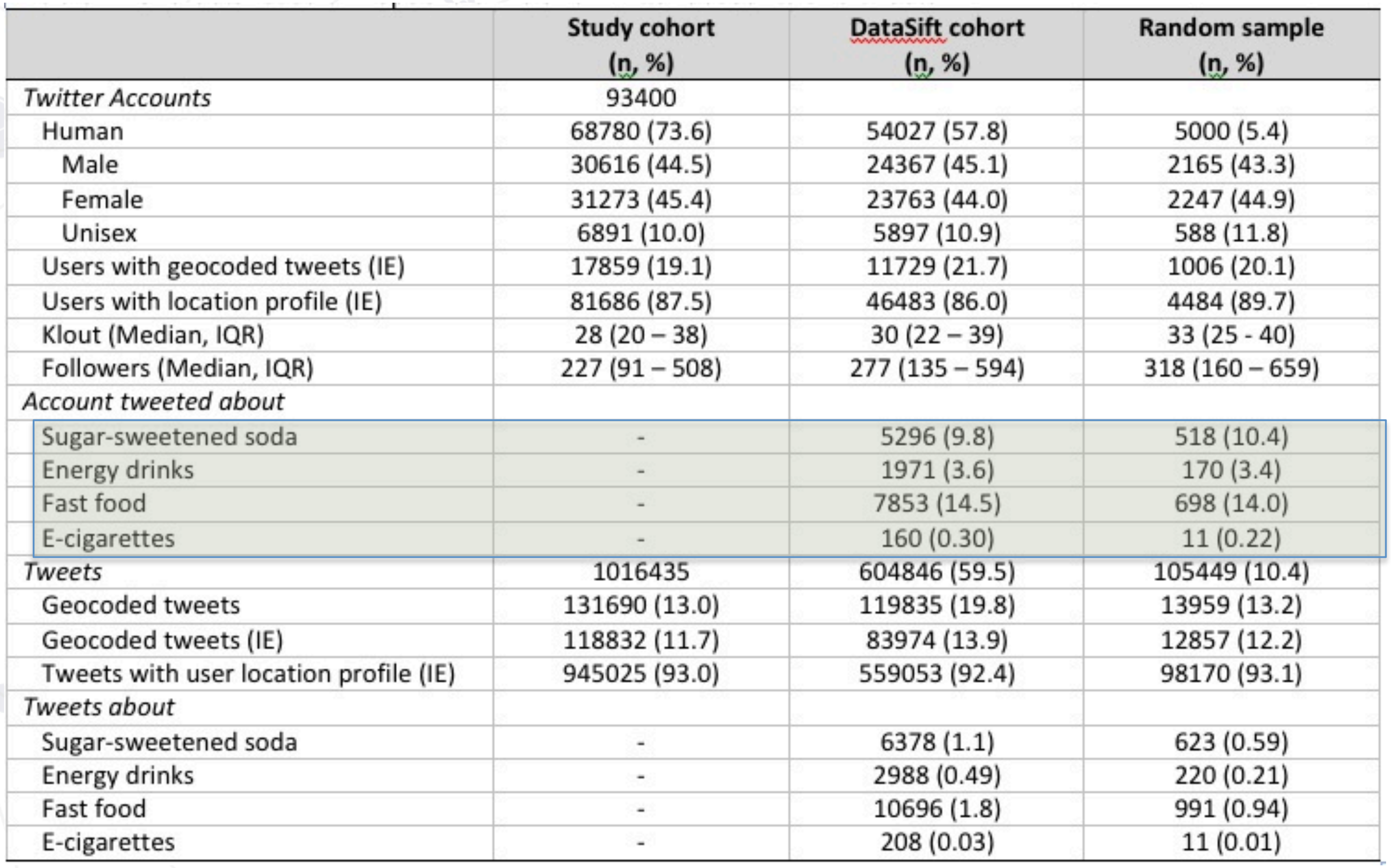
Centre for
Data Analytics

NZ

Insight

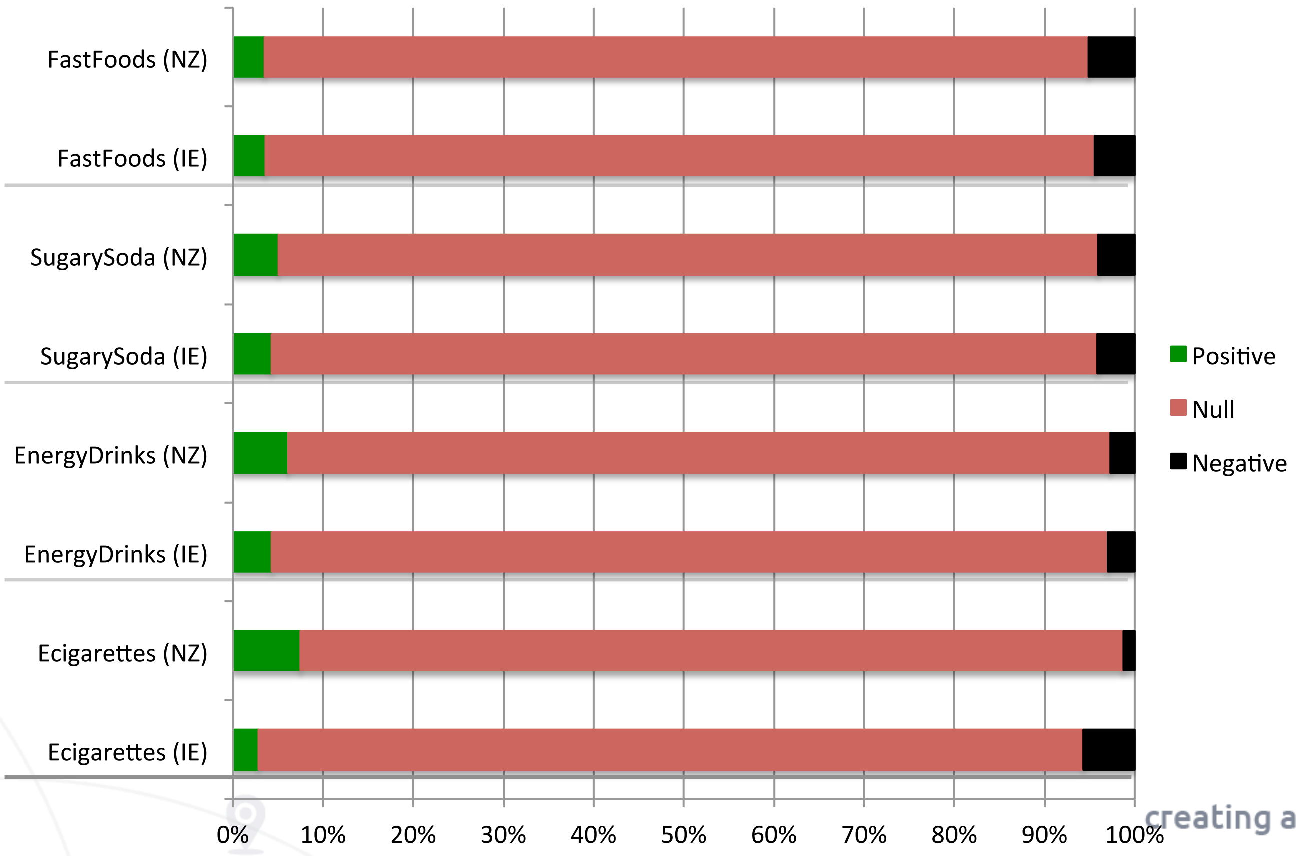


	Study cohort (n, %)	DataSift cohort (n, %)	Random sample (n, %)
Twitter Accounts	60784		
Human	41887 (68.9)	16259 (38.8)	5000 (11.9)
Male	18179 (43.4)	7304 (44.9)	2275 (45.5)
Female	18799 (44.9)	6902 (42.5)	2096 (41.9)
Unisex	4912 (11.7)	2053 (12.6)	629 (12.6)
Accounts with geocoded tweets (NZ)	9039 (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)



	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)

DataSift's Sentiment Analysis on 5,000 ?



- … and then DataSift and Twitter parted company and we were left high, and dry !

What have we achieved ?

- Assembled a cohort of Twitter accounts, NZ and IRL, using Twitter profiles, specific to 4 topics of interest
- Accumulated tweets from this cohort, and randomly sampled 5,000 of these accounts creates a cohort representative of the larger cohort, both with respect to the account and the tweet characteristics
- All other investigations have used a sample of tweets, rather than sampled accounts
- For *fast foods*, *sugar-sweetened beverages* and *energy drinks* we don't need the firehose, and by building cohorts of accounts, we bypass bots and malicious or malevolent "gaming" of sentiment and volume-based analysis
- This is a "Feasibility Study", next step(s) are to track these accounts' postings over time, replicating what the pollsters do in conventonal polling