Investigating Public Health Surveillance using Twitter

Antonio Jimeno Yepes[♢♠], **Andrew MacKinlay**[♢♠], **Bo Han**[♢]
[♢] IBM Research Australia, Melbourne, VIC, Australia
[♠] Dept. of Computing and Information Systems, University of Melbourne, Australia
{antonio.jimeno,admackin,bohan.ibm}@au1.ibm.com

Abstract

Microblog services such as Twitter are an attractive source of data for public health surveillance, as they avoid the legal and technical obstacles to accessing the more obvious and targeted sources of health information. Only a tiny fraction of tweets may contain useful public health information but in Twitter this is offset by the sheer volume of tweets posted. We present a system which can identify medical named entities in a real-time stream of Twitter posts and determine their geographic locations, as well as preliminary experiments in using this information for health surveillance purposes.

1 Introduction

Public health surveillance (Nsubuga et al., 2006) is the systematic collection, analysis and monitoring of population health for the public good using a variety of tools. For instance, syndromic surveillance (monitoring for symptoms as signatures of diseases) can be used for tracking and early detection of infectious diseases to flag potential outbreaks, assist in disease modelling, or detect cases of biological terrorism. Meanwhile, pharmacovigilance (WHO and others, 2002) can be used to detect adverse effects associated with pharmaceutical products, while statistics on population health and wellbeing can inform governmental health policy. However, to be effective, these applications require large volumes of real-world data on health statistics (such as from hospital records), which are in most cases difficult to access because of privacy regulations and technical challenges.

The proliferation of social media might enable legitimate large scale collection of health information. Users of forums (e.g., PatientsLikeMe) and microblogs (e.g., Twitter), which we focus on here, post health-related messages with varying levels of frequency. These might cover diseases they have, symptoms they have experienced or drugs they have taken. Twitter may have a large enough volume of data to partially make up for its lack of a health-specific focus. Some judiciously-used data is better than no data at all which is often all that can be obtained from health-specific sources. Such information can be leveraged in analytics to provide insights on public health, e.g., for drug safety (Sarker et al., 2015). However, it is still unclear how large a contribution social media could make to population health surveillance.

In this paper, we perform analysis of health related Twitter data for public health surveillance. The large volume of data in Twitter (approximately 5000 posts per second) is the reason it is useful for such tasks, but each of these posts must be examined (in real-time for practical applications) to determine whether is it relevant, and if so, stored for subsequent analysis. Here, we consider a relevant post to be one containing medical named entities, as identified by an in-domain named-entity tagger (Jimeno Yepes et al., 2015) which we run over our entire data-set after applying some pre-filtering heuristics. A second challenge with Twitter is that location information is scarce, with only around 2% of messages containing reliable geographic coordinates (Cheng et al., 2010). Location information is needed, for instance, in syndromic surveillance to identify the possible location of an outbreak. We handle this by adapting and tuning an existing geotagger to augment the tweets with automatically-determined geographic information (Han et al., 2013). We

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 164–170,
Beijing, China, July 30, 2015.

then analyse the data, by examining the trend of geolocated medical entities in different regions, presenting commonly discussed medical entities in different categories, and identifying salient medical entities and common topics for a given medical entity. Our results show promising outcomes of utilising Twitter data in health surveillance applications and also raise some limitations of using this data. Overall, the contributions of this paper are twofold: (1) it helps us to understand to what extent Twitter data supports public health surveillance and (2) it provides pilot results that indicate future directions to explore when utilising Twitter data for public health.

2 Related Work

Several sources of data have been previously considered for public health surveillance. Biosurveillance has been usually achieved by monitoring emergency department notes (Espino et al., 2004). The data is reliably sourced, however, there are severe issues in processing time and data aggregations when the data is collected from several departments in various forms and with different time latencies. In addition, access to these sensitive electronic health records is also restricted by privacy issues.

Search engine query logs are an abundant source of data for the organisations which own the search engines, and have been exploited in the health realm. Google[1] (Carneiro and Mylonakis, 2009) finds a spatio-temporal correlation between flu-related queries and data from the United States Centers for Disease Control (CDC). Similarly, Yom-Tov and Gabrilovich (2013) have used Yahoo search data to identify adverse-drug reactions. However, since the search logs are not publicly accessible, these methods are only viable for the companies which own the search log data.

An alternative approach is to monitor information from news data. Collier et al. (2008) identified health rumours and compared them to CDC data, however this might be less successful for real time monitoring and less public disease outbreaks, because only large outbreaks of diseases are newsworthy, and they will have some time lag. For health information of individuals, it is more likely to appear in search logs or medical forums (Segura-Bedmar et al., 2014; Metke-Jimenez et al., 2014; Cameron et al., 2013).

Twitter data has also been considered to identify trends in the 2009 swine flu outbreak in the UK that correlated with official data (Lampos and Cristianini, 2010) and to track alcohol consumption (Kershaw et al., 2014) using geolocated tweet data. Some initial work on exploring health topics in Twitter has been previously done (Paul and Dredze, 2011; Paul and Dredze, 2012; Prier et al., 2011; Signorini et al., 2011), showing the presence of health-related information. These systems typically rely on the Twitter API data with location information.

While there has been some work on medical text mining in social media (e.g., identification of relevant tweets for adverse drug events (Nikfarjam et al., 2015)), a critical assessment of performance of current text mining technology has not been performed. In this work, we have taken a closer look into Twitter data for public health surveillance.

3 Methods

Our pipeline for processing and analysing the Twitter stream is represented in Figure 1. Medical named entities are identified in tweets and those tweets are then geotagged if they do not contain accurate GPS labels. From the large volume of source Twitter data, this yields a much smaller number of tweets containing of medical named entities along with geographical information. This smaller data set is then stored in a MongoDB[2] document database for querying and filtering.

3.1 Micromed: medical NER for Twitter

We have developed a medical named entity recogniser, named `Micromed` (Jimeno Yepes et al., 2015), which uses supervised learning to recognise three types of entities: diseases, symptoms and pharmacological substances.[3] It uses a linear-chain CRF (condi-

[1] Google Flu Trends: `http://www.google.org/flutrends`

[2] `https://www.mongodb.org`

[3] For performance reasons the CRF implementation used here was different to the original system and no POS-based features were used, resulting in a roughly

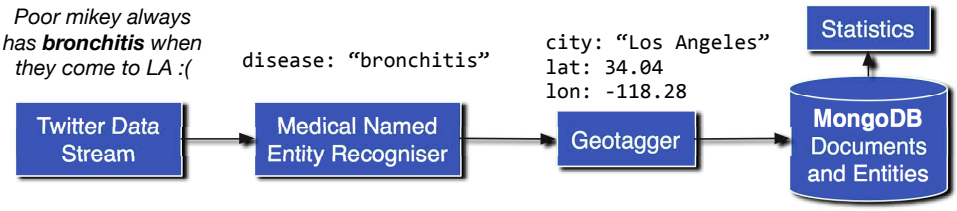

Figure 1: Annotation pipeline

tional random field) (Lafferty et al., 2001), and is trained on a publicly available[4] set of 1300 tweets which have been manually annotated with relevant medical entities. The three entity types are correspond with entries in the Unified Medical Language System (UMLS) (Bodenreider, 2004) Semantic types – specifically *T047 (Diseases or Syndrome)* for diseases, *T184 (Sign or Symptom)* for symptoms and *T121 (Pharmacologic Substance)* for pharmacologic substances. Table 1 shows the performance of `Micromed` on our annotated set for exact matching of the boundaries of the entities, which outperforms systems like MetaMap (Aronson and Lang, 2010) or Stanford NER (Finkel et al., 2005). A comparison is available in (Jimeno Yepes et al., 2015).

Entity Type	Precision	Recall	F1
Disease	0.7987	0.5020	0.6165
Pharm.Subs.	0.8142	0.3948	0.5318
Symptom	0.7193	0.6028	0.6559

Table 1: `Micromed` performance evaluated using 13-fold cross-validation

3.2 Geotagger

To obtain geolocation information for the vast majority of tweets, we adapted and tuned an off-the-shelf geotagger `LIW-META` (Han et al., 2013). `LIW-META` leverages location indicative words to infer geolocations for tweets which lack GPS labels. It applies various feature selection methods to extract words associated with particular locations. Both explicit gazetted terms (such as city and country names) and implicit location-indicative words (such as local landmarks, sport teams and dialectal terms) are extracted and used in modelling taggers. Additionally, it also exploits user profile data such as user-declared locations and time zone information in a stacking framework to enhance the prediction accuracy (Han et al., 2014).

3.3 Twitter data set

We used all of the tweets from 2014[5] obtained from GNIP Decahose,[6] which provides 10% of tweets randomly selected from Twitter. In a pre-filtering step, we remove the 33.5% of posts marked as retweets (which are less interesting for our use cases) and the 70.5% that were marked as non-English (which our tagger is not designed for). The remaining tweets (23.3% of the tweets in the GNIP decahose overall) are processed using the pipeline in Figure 1 and stored if a medical entity was found.

4 Results

In this section, we explore the tweets that contain medical entities to understand what information it might be possible to extract from them. We first have a closer look at the medical entities extracted by `Micromed` and the extended coverage obtained from the geotagger. The coverage of `LIW-META` is further displayed showing statistics for several large cities.

4.1 Medical entities

The statistics for the number of tweets at each phase of the pipeline are summarised in Table 2. 27 million tweets had at least one medical entity, corresponding to 1.0 tweets per second (83k tweets per day) from the GNIP decahose, which would correspond to 10 tweets per second on the full live Twitter stream. Unsurprisingly, this proportion containing medical information is only a small fraction (around 0.2%) of the tweets in the Decahose stream.

1.5% drop in F-score

[4] `https://github.com/IBMMRL/medinfo2015`

[5] Apart from a gap from February 25 to March 22 in our dataset

[6] `https://gnip.com/sources/twitter`

Stage	Total	Per day	Kept
Decahose	$12{,}000\times10^6$	$36{,}254\times10^3$	–
Pre-filtered	$2{,}800\times10^6$	$8{,}459\times10^3$	23.3%
Medical	28×10^6	83×10^3	1%

Table 2: Statistics for tweet numbers initially, pre-filtered (removing non-En and retweets) and discarding tweets without medical entities

We have listed the most frequent annotated entities for each type in Table 3. Some entries are not particularly surprising: substances like *marijuana* or *caffeine*) and symptoms like *tired* or *hungry* are likely to be reflective of the frequency of people using or experiencing these. However diseases such as *heart attack* are less likely to indicate actually occurrences of that disease. Since the volume of tweets with medical entities makes it difficult to interpret the context of the entities mentioned, we have used the `MALLET` (McCallum, 2002) implementation of topic modelling (Blei et al., 2003) to group the tweets by topic.

Table 4 shows 5 topics for *heart attack*. Except for topic 3, related to the memory of people who suffered the disease, in most cases the use of the term seems to have a figurative connotation related to excitement, which indicates that additional work is required to identify tweets to discard figurative terms (and possibly historical events).

Table 5 shows the topics for *marijuana*. In most cases, the topics are related to legalisation of marijuana in the USA. Whether this has a correlation with actual usage rates, and thus potential impact in public policy for example, requires further investigation.

Topics for entity *tired* are shown in Table 6. In some topics, *tired* seems to be used figuratively to express being bored or impatient. Again, the ability to accurately identify figurative uses of terms could be valuable.

4.2 Geolocation

Location information for each tweet is needed, for instance, to identify the location of an outbreak. Overall, 4.8% of tweets come with GPS labels in our English GNIP collection. Not all tweets are equally predictable so we have calibrated `LIW-META` by selectively choosing reliable prediction indicators. We tested whether the overall prediction is more reliable when its sub-predictions agree with each other and we found that the overall prediction is more accurate when it agrees with predictions based on user declared locations. This calibrated setting achieves 0.938 precision and 0.214 recall using all geotagged tweet data for evaluation. Our Twitter set offers 0.6 million GPS-labelled tweets while Twitter + `LIW-META` generates 8.9 million tagging results.

4.3 Geotagged tweets with medical entities

The subset of tweets containing medical entities have been enhanced with location information from the geotagger. Figure 2 shows the number of tweets for three large cities (New York City, London and Chicago) during part of the first half of 2014. The geotagger used here significantly increases the number of health-related tweets that can be identified belonging to these large cities.

5 Discussion

From the large number of tweets being posted every second, just a small fraction of 0.2% (10 per second) contain medical terms. Despite this, a large number of tweets still provide relevant health information.

Twitter poses additional challenges compared to traditional NLP in medical literature and clinical text. Many tweets lack standard grammatical structure or possess abbreviations and misspellings (Baldwin et al., 2013). The use of figurative language in Twitter may be more frequent than other domains (it is clearly very common in our data for many of the frequent symptoms and diseases), although it is particularly important to disambiguate this here for most of the proposed used cases. However there are cases in which the context of the entity makes a medical entity seem legitimate to the tagger (e.g. *heart attack*), so additional filtering might be required.

6 Conclusions

This paper augments in-domain NLP tools to extract and analyse medical information in Twitter. We find the overall proportion of tweets with medical entities is small, nonetheless, we are able to harvest a respectable num-

Disease	Frequency	Pharm. Sub.	Frequency	Symptom	Frequency
heart attack	374810	marijuana	379838	tired	5075812
cancer	268988	caffeine	114526	hungry	2885491
diabetes	175992	cannabis	100233	pain	1724314
stroke	161549	heroin	93723	headache	980699
aids	131792	alcohol	64957	stress	947341

Table 3: Most frequent entities annotated by `Micromed` per entity type.

1	love, guy, put, feel, direction, knew, mtvhottest, heart, https, line
2	phone, mini, dropped, alarm, drop, screen, show, fire, case, find
3	dad, died, find, massive, ago, couldn, told, years, today, days
4	heart, attack, read, seconds, reading, part, summer, book, words, min
5	eat, food, eating, bacon, burger, plate, cheese, grill, pizza, ate

Table 4: Top 5 topics for entity *heart attack*

1	http, tv, legalization, live, job, reporter, fight, vending, machine, quit
2	arrested, possession, police, jail, texas, charges, arrest, son, man, officer
3	tax, million, weed, legalizeit, year, shouldbelegal, sales, revenue, taxes, billion
4	legalized, states, bowl, super, legal, legalize, seattle, united, teams, recreational
5	alcohol, marijuana, dangerous, california, worse, difference, safe, decide, tobacco, human

Table 5: Top 5 topics for entity *marijuana*

1	tired, haha, damn, xd, ah, la, tmr, meh, hmm, uh
2	tired, omg, damn, stand, understatement, joke, soooo, social, omfg, soooooo
3	tired, anymore, isn, point, word, part, fight, basically, helping, state
4	don, wanna, feel, sleep, understand, worry, honestly, numb, aware, bothered
5	tired, soo, sleep, damn, gosh, fucken, darn, crabby, frick, aswell

Table 6: Top 5 topics for entity *tired*

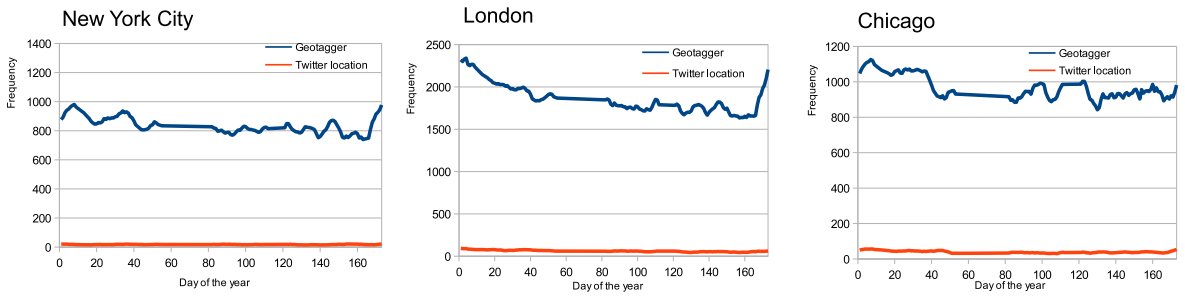

Figure 2: Seven day rolling average of tweets with medical entities count per day for New York city, London and Chicago for January–June 2014

ber of refined medical entities due to the sheer volumes of Twitter data. We extract frequent medical entities in three pre-defined categories, highlight the collocations with entities and investigate topics where an entity is mentioned. By further assigning entities with geographical locations, we can obtain better local medical trend signals which makes public surveillance more plausible. Overall, we have found evidence for the plausibility of public health surveillance using Twitter, although there is much scope to expand on our data analysis in the future.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. 2013. Predose: A semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics*, 46(6):985–997.

Herman Anthony Carneiro and Eleftherios Mylonakis. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010)*, pages 759–768, Toronto, Canada.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Jeremy U Espino, Michael M Wagner, Fu-Chang Tsui, H Su, Robert T Olszewski, Z Lie, Wendy Chapman, Xiaoming Zeng, Lili Ma, Z Lu, et al. 2004. The rods open source project: removing a barrier to syndromic surveillance. *Medinfo*, 2004:1192–6.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 7–12, Sofia, Bulgaria, August.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal Artificial Intelligence Research (JAIR)*, 49:451–500.

Antonio Jimeno Yepes, Andrew MacKinlay, Bo Han, and Qiang Chen. 2015. Identifying Diseases, Drugs and Symptoms in Twitter. *MEDINFO*.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2014. Towards tracking and analysing regional alcohol consumption patterns in the uk through the use of social media. In *Proceedings of the 2014 ACM conference on Web science*, pages 220–228. ACM.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit.

Alejandro Metke-Jimenez, Sarvnaz Karimi, and Cecile Paris. 2014. Evaluation of text-processing algorithms for adverse drug event extraction from social media. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 15–20. ACM.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, page ocu041.

Peter Nsubuga, Mark E. White, Stephen B. Thacker, Mark A. Anderson, Stephen B. Blount, Claire V. Broome, Tom M. Chiller, Victoria Espitia, Rubina Imtiaz, Dan Sosin, Donna F. Stroup, Robert V. Tauxe, Maya Vijayaraghavan, and Murray Trostle. 2006. Public health surveillance: a tool for targeting and monitoring interventions. In *Disease Control Priorities in Developing Countries. 2nd edition.* World Bank.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, pages 265–272.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Kyle W Prier, Matthew S Smith, Christophe Giraud-Carrier, and Carl L Hanson. 2011. Identifying health-related topics on twitter. In *Social computing, behavioral-cultural modeling and prediction*, pages 18–25. Springer.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, 54:202–212.

Isabel Segura-Bedmar, Santiago de la Pena, and Paloma Martınez. 2014. Extracting drug indications and adverse drug reactions from spanish health social media. *ACL 2014*, page 98.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

WHO et al. 2002. The importance of pharmacovigilance. *Geneva: World Health Organization*.

Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6).