Text Mining from Social Media for Public Health Applications

Joana M. Barros
Insight Centre for Data Analytics - NUI Galway
IDA Business Park, Lower Dangan
Galway, Ireland
joana.barros@insight-centre.org

ABSTRACT

Public Health is crucial to manage and monitor threats to the health of the population. In recent years, Twitter has been successfully applied to monitor diseases through its ability to provide near real-time data and proved to be an asset to the domain. This research aims to further explore capabilities of Twitter in the disease surveillance field by focusing on its geolocation feature and health mentions, identifiable through disease-specific language patterns present in Twitter messages.

CCS CONCEPTS

•Applied computing → Life and medical sciences; Health informatics;

KEYWORDS

Public Health; Disease Surveillance; Social Media; Twitter

ACM Reference format:
Joana M. Barros. 2017. Text Mining from Social Media for Public Health Applications. In *Proceedings of DH '17, July 2–5, 2017, London, United Kingdom,* , 2 pages.
DOI: http://dx.doi.org/10.1145/3079452.3079475

1 PROJECT OUTLINE

1.1 Introduction

Over the course of their evolution, humans have continuously dealt with agents and ailments that dictate their mortality. Advancements in the fields of medicine, biology and microbiology were vital to improve the world's population health leading to a longer life expectancy. Nowadays, Public Health is crucial to detect, avoid, deal and monitor threats to the health of the population and has benefited from the use of surveillance which permits a systematic collection and analysis of health information [5]. Traditionally, surveillance relies on a network of health facilities and laboratories which report to official health entities. Despite the data high quality, this system can be costly leading to reporting delays and affecting the rapid detection of disease outbreaks. Syndromic surveillance deals with these issues by using sources available before a diagnosis is confirmed (e.g. over the counter drug sales). This type of surveillance is based on the assumption that an outbreak would manifest itself as an anomaly in behaviour [7]. To adequately prepare for an outbreak, governmental and local health entities require early warnings to deploy adequate measures. This has become a key issue for Public Health and it instigated the application of new sources of valuable health information. Modern sources of data (e.g. search engine queries [4] and on-line news [1, 3]) can provide near real-time, government independent outbreak information in various formats. Recently, special attention has been given to social network sites such as Twitter which has been applied to monitor disease awareness [13] and for disease surveillance [11], suggesting its possible applications for evaluating the health state of a population.

1.2 Motivation

Monitoring the health state of a population has become a fundamental issue. Globalisation heightens the difficulties in the application of measures to contain outbreaks thus, early detection of outbreaks has proven vital for a swift intervention by the health authorities, enabling fewer negative impacts in the population [5]. This research is based on the hypothesis that large-scale social media data can provide new insights about the health state and mobility patterns of the population. My main hypothesis is further extended with the following:

(1) Twitter can be used as a mean to identify travel patterns, which may provide an explanation for changes in the prevalence of a disease [6, 12].
(2) Variations in disease frequency can be monitored trough mentions in user generated messages [2, 8].
(3) Health related messages can provide information concerning disease-specific language patterns. This would be used as an unsupervised method to detect accurate disease mentions.

The final goal of this research will be the development of an algorithm capable of continuously monitor the frequency of a wide range of diseases while taking into account the proximity to travel and disease prone hubs (i.e airports and hospitals). This would be a suitable addition to syndromic surveillance and, in addition to its benefits as an early detection tool for infectious outbreaks, it could also be applied to identify raises in non-infectious diseases frequency.

2 ONGOING APPROACH AND RESULTS

2.1 Data Collection and Preprocessing

In the initial stages of this research, our case study is the clinical terms from the Systematized Nomenclature of Medicine - Clinical Terms, accessed trough Bioportal [14]. This source was chosen due to its widespread use, as well as its comprehensive and precise nature in describing clinical terms. The final collection amounted to 2468 disease/clinical terms, including synonyms.

DH '17, July 2–5, 2017, London, United Kingdom
 ACM ISBN 978-1-4503-5249-9/17/07.
DOI: http://dx.doi.org/10.1145/3079452.3079475

Tweet:
I got **DISEASE** at the State Fair

Patterns:
POS-1 **DISEASE** POS1
POS-1 **DISEASE** POS1 POS2
POS-2 POS-1 **DISEASE** POS1
POS-2 POS-1 **DISEASE** POS1 POS2

Figure 1: Patterns are created around the disease term. When it is not possible to have an equal number of POS on each side, the creation of new patterns is ceased. For this example four patterns were created.

Table 1: Disease Terms Search

Collection	Number of Tweets	Highest Frequency
Hospital	48	26 (Mental disorder)
Airport	187	41 (Mental disorder)
Geolocated	8694	1853 (Mental disorder)

The Twitter data for this stage comprehends a continuous collection of 11,163,218 tweets gathered from October-November 2016, using the Twitter API. This data set was then filtered based on the tweets proximity to airports ("Airport" collection) [10] and 39 manually selected hospitals ("Hospital" collection) totalling 177,238 and 32,106 messages respectively. The remaining tweets were labelled as belonging to the "Geolocated" collection.

2.2 Language Patterns

As an early attempt to identify language patterns each tweet was divided into tokens and a Part-Of-Speech (POS) tag [9] was assigned to each token. It was decided to focus on POS tags due to their ability to provide a general grammatical tag based on a word definition and its context. To produce the POS patterns a rule-based approach, exemplified in Figure 1, was followed.

The preliminary results of this approach, represented in Table 1, suggest that potential health mentions are scarcely discussed on Twitter. This was suggested by the low amount of tweets selected after the search for disease terms mentions. Also, mentions regarding the "mental disorder" super-class returned the highest frequency with the majority of the disease terms corresponding to "anxiety disorder" through the sub-class synonym "nightmare".

Regarding the POS patterns, the results suggest a constant presence of singular proper nouns preceding and succeeding the disease mention. An exception to this are the patterns where an adjective and a preposition or subordinating conjunction precede the disease term, which occurs in the "Hospital" and "Geolocated" collections, and patterns where a coordinating conjunction occurs after the disease term, present in the "Airport" collection. However, these results come with a caveat as the majority of the tweets did not apply the disease terms in a clinical sense.

3 FUTURE WORK

The detection of health mentions in social media data poses several challenges. At this research stage, the ambiguity associated with the disease terms was confirmed; although present in Twitter messages, the terms were seldom used in the clinical sense. This hindered the identification of linguistic patterns as the ones identified may solely relate with the nature of tweets and not with the diseases themselves. To address this, I plan to implement distributional semantics notions to enrich the language pattern identification. In addition, a larger corpus of data will provide more context associated with the disease terms and, for example, the use of syntactic dependencies will aid in the terms semantic interpretation. With this in consideration, the immediate purpose of this stage is to identify the limits of unsupervised techniques for semantic disambiguation in Twitter messages, when applied to the health domain.

4 COMPETING INTERESTS

The authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

This research has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

REFERENCES

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

[2] Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*. ACM, 115–122.

[3] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* 15, 2 (2008), 150–157.

[4] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[5] David M Hartley, Noele P Nelson, RR Arthur, P Barboza, Nigel Collier, Nigel Lightfoot, JP Linge, E Goot, A Mawudeku, LC Madoff, and others. 2013. An overview of Internet biosurveillance. *Clinical Microbiology and Infection* 19, 11 (2013), 1006–1013.

[6] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41, 3 (2014), 260–271.

[7] Kirsty Hope, David N Durrheim, Edouard Tursan dfiEspaignet, and Craig Dalton. 2006. Syndromic surveillance: is it a useful tool for local outbreak detection? *Journal of Epidemiology and Community Health* 60, 5 (2006), 374–374.

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

[9] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1 (ETMTNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 63–70. DOI:http://dx.doi.org/10.3115/1118108.1118117

[10] OurAirports. 2016. Open data downloads. http://ourairports.com/data/. (2016). (Accessed on 13/09/2016).

[11] Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from Twitter. *Health* 11 (2012), 16–6.

[12] Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Modeling Spread of Disease from Social Interactions.. In *ICWSM*.

[13] Michael Smith, David A. Broniatowski, Michael J. Paul, and Mark Dredze. 2016. Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

[14] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research* 39, suppl 2 (2011), W541–W545.