How reliable is Twitter for influenza surveillance?

Servan Grüninger

July 23, 2016

Contents

1 Administrative Information

General Reinhard Furrer (RF) & Marcel Salathé (MS) will jointly supervise project. RH will be primary supervisor regarding statistical questions; MS will be primary supervisor regarding machine learning questions.

EPFL Sharada Mohanty (SM) & Gianrocco Lazzari (GL) are PhD students in MS' group with knowledge in data management. They act as first contact persons regarding data management and processing question.

UZH Reinhard Furrer will assign a PhD student who acts as first contact person for statistical questions arising during project.

Abbreviations

- Reinhard Furrer = RF
- Marcel Salathé = MS
- Servan Grüninger = SG
- Sharada Mohanty = SM
- Gianrocco Lazzari = GL

2 Current & Achieved Goals

2.1 Support Vector Machines & Naive Bayes Classifiers

Added: 2016.03.20
Task: Continue reading into theory behind support vector machines and naive bayes classifiers (see also progress report 3.3)
Status: In process

2.2 C4.5 & meta-classifiers

Added: 2016.03.20
Task: Get acquainted with principles behind C4.5 algorithm and meta-classifiers (see also progress report 3.3)
Status: In process

2.3 Explorative Data Analysis

Added: 2016.02.04
I will have a look at the algorithm itself; get acquainted with it; read, explore and visualise some data with it.
Task: Get algorithm to work on my local environment; test it out with small data sets
Status: not started

2.4 Define Fine-Grained Goals - Done

Added: 2016.02.04
In order to plan the next steps, we need fine-grained goals concerning the master project
Task: Define fine-grained goals based on general goals defined during last meeting (see 3.3)
Status: Done (see 3.3)
Finished: 2016.03.20

2.5 Define General Goals - Done

In order to narrow down the direction of the project, we need general goals concerning the master thesis
Added: 2016.02.04
Task: Define general goals
Status: Done (see 3.4)
Finished: 2016.02.04

2.6 Theoretical Background - Done

Added: 2016.02.04
In a first step, SG shall get acquainted with the theoretical prerequisites to understand and use the algorithm.
Task: Read into literature cited in [Bodnar et al., 2014] and related work; generate an overview over relevant topics which I have to refresh or get acquainted with.
Status: Done (see 3.4)
Finished: 2016.03.20

3 Updates & Meetings

3.1 2016.07.20 - Progress Report

3.1.1 Exploratory Analysis of Twitter Data

I used GL's summary and script for a first analysis of the data see ("Masterarbeit/TwitterData/tweets_from_todd" for script and summary

Questions/Key observations:

What does "sick" label mean in this context? I assume that the label "sick" in the data set shows the result of Todd's code. Need to double-check with MS & GL

test test

3.1.2 Exploratory Analysis of Twitter Data - Code

3.2 2016.05.23 - Progress Report

3.2.1 Remarks on Todd Bodnar's Dissertation

Todd decided to create six subsets from twitter data set

- entire data set grouped by each week
- tweets containing flu keywords ("flu", "cough", "fever", "headache", "head ache")
- tweets containing zombie keywords ("zombie", "zed", "undead", "living dead")
- same three subsets above, but divided based on which region tweet came from

The idea behind using the zombie keywords was the assumption that they would be unrelated to the flu. However, this is most likely not the case, since some people might tweet about "feeling like a zombie" or using #zombie, #undead when talking about having the flu. Urban Dictionary even knows the expression "Zombie flu" http://www.urbandictionary.com/define.php?term=zombie+flu ; https://twitter.com/SnKTyphooon/status/726593663095177216

1000 most common words in each subset were used as list of keywords for the models (keyword trends were measured by their frequency due to fluctuations)

Models used:

- univariate logit: $logit(CDCRate) = \beta_0 + \beta_1 logit(x) + \epsilon$; x is the fraction of tweets containing at least one ILI keyword
- multivariate logit: $logit(CDCRate) = \beta_0 + \sum_{i=1}^{n} \beta_i logit(x_i) + \epsilon$; x_i is the frequency of the i^{th} keyword
- select best keywords (only use that keyword for regression fitting)
- SVM regression

Data sample:
104 individuals with influenza, 122 individuals without influenza. How was absence of influenza assessed? Could individuals have had influenza without having gone to the doctor?
Twitter handle only available for 119 individuals, of which 15 were discarded; two additional discarded because they tweeted too often. Totally 37599 tweets from seed accounts and 30950958 tweets from 913082 connected accounts
Question: When did tweet collection start? I.e. did it start *after* or *before* diagnosis? And did users know that they participated in the study?

3.3 2016.03.20 - Progress Report

3.3.1 Theory

I looked into the models mentioned in [Bodnar et al., 2014] and started refreshing and/or getting acquainted with them. Current status:

- random forests: basic knowledge present, I just refreshed some key concepts
- logistic regrssion: basic knowledge present, I just refreshed some key concepts
- support vector machines: no previous knowledge; I started reading into theory (starting from [Burges, 1998], [Hearst et al., 1998] and online resources)
- naive bayes classifiers: no previous knowledge; I started reading into theory (starting from [McCallum et al., 1998], [Lewis and Ringuette, 1994], [Rennie et al., 2003] and online resources) enditemize
- C4.5 classifier: no previous knowledge; did not start reading yet
- metaclassifiers: no previous knowledge; did not start reading yet

3.3.2 Courses

I'm taking a total of two (three) courses at EPFL and one course at UZH - all revolving around topics/subjects I'll need in my master thesis

- Parallel and High-Performance Computing (5 credit points, EPFL)
- Convex Optimization And Applications (4 credit points, EPFL)
- Computational Linear Algebra (4 credit points, EPFL) - I will probably drop this one, however
- Advanced R progrmaming (3 credit points, UZH)

3.3.3 Questions, Remarks & Goals

The goals below are a first suggestion from my side. The list is not supposed to be complete and the single goals are open for changes and amendments

How representative are geotagged tweets? From the report forwarded to me by GL, we can see that the highest peak in tweets/day occurs in autumn 2013 - without apparent explanation. Also, in [Bodnar and Salathé, 2013] it is also mentioned that the restriction of the dataset to geotagged tweets could introduce biases. Finally, not every tweet that is sent by the same user is geotagged. Hence, we would need to find out how complete each user's tweet history is. If geotags are only added to tweets fo specific purposes, our dataset will be heavily biased.

Is there any possibility to compare our geo-tagged tweets with a representative sample from the general twitter population? For example, I would like to know whether the maximum in tweets/day occured also in autumn 2013 or not. Intuitively, I would have believed that the Ebola craze in 2014 would have created many more sickness-related tweets than any other event in the recent years. At least when it comes to media-reporting, the Ebola epidemic was by far the most talked-about disease-related event in the past 15 years (see mountainsOutOfMolehills), so I would have assumed that we could observe a similar pattern on Twitter

Goal: Compare tweets from geotagged dataset with a representative sample of the "general twitter" population. Find out whether geotagged tweets have specific biases. Also, find out whether geotagged tweets of one user consist of all the tweets which were sent by said user

How representative is the dataset used to build the algorithm? From [Bodnar et al., 2014] we learn that the data set used to build the algorithm consisted of totally 104 twitter accounts generating a total of 37'599 tweets. Out of this sample, 35 users fell sick during the study period and generated a total of 1609 tweets in the month in which they were sick. Furthermore, all twitter users stemmed from the same state (Pennsylvania) and belonged to approximately the same socio-economic group (young students of the Pennsylvania State University). Hence, one would assume that their tweeting behaviour is different from that of the average twitter user. Hence, we need to test the performance of the algorithm for different cities and states and compare the results with reliable epidemiological data.

Goal: Apply algorithm to small datasets from other area. Using cities within the Pennsylvania, neighbouring states as well as randomly chosen states and cities from all over the US in order to test the algorithm and see whether results make sense (aka hold up against a comparison with CDC data)

What is the temporal resolution of the algorithm? Due to privacy concerns, the authors of [Bodnar et al., 2014] only knew in which month an indiviual was diagnosed with influenza. Hence, this might heavily reduce the temporal resolution of the algorithm. Hence, we should assess how the algorithm performs for assessing the influenza infection rates in time frames which are shorter than a month.

Goal: Find out the temporal resolution of the algorithm by comparing algorithm prediction with CDC data

How do we label the training set when using large datasets? Due to the small sample size, the tweets in Bodnar et al. [2014] were labelled manually in order to identify tweets that directly talk about influenza and those that don't. For larger data sets, this won't be possible anymore. Hence, we should either resort to Amazon mechanical turks or try to implement code

that can distinguish between influenza-related and -unrelated tweets. Possible candidates are described in [Paul and Dredze, 2011] (Ailment Topic Aspect Model), Culotta [2010], and (probably most relevant for our purpose) [Lamb et al., 2013].
Goal: Implement algorithm that identify between tweets that are flu-related, concern the tweeter himself (and not one of his relatives or friends) and talk about an infection (as opposed to general awareness about the flu)

Can the performance of the algorithm be improved by incorporating prior knowledge? According to [Goel et al., 2010] the incorporation of prior information can help to improve the performance of an algorithmed destined to identify twitter patterns linked to disease outbreaks. Prior knowledge can either be used to refine keyword searches ([Lamb et al., 2013, Paul and Dredze, 2011]) or to improve the prediction model itself by incorporating results from other web data-based prediction models - like ”google flu trends” ([Ginsberg et al., 2009], ”Sickweather” or ”FluNearYou” Butler [2013]) - and/or results from autoregressive models based on CDC data or data from local authorities (like the Emergeency Department Fever and Respiratory Complaint Surveillance in New York City [Olson et al., 2007].
Goal: Test whether the incorporation of prior information enhances the performance of the algorithm

Optional: Can we incorporate Spanish tweets as well? This would give us more data points to train and expand the model. Also, it will tell us if the algorithm holds up for twitter users speaking a different language but living in the same country/region. The people behind Google Flu Trends expanded their algorithm to Mexico, so we might copy some ideas from them Google Faculty Summit 2009: Google Flu Trends
Goal: Incorporate Spanish tweets

3.3.4 Next Steps (by Sunday, April 10th)

- start with explorative data analysis (see goal 2.3)
- continue theory review (i.e. goals 2.1 and 2.2)
- finalize concrete project plan (for UZH Mastervereinbarung)

3.4 2016.02.04 - Skype Meeting

We discussed the tentative Master project plan in order to narrow down the possible direction of the thesis. The following three goals were defined:

Extension of Algorithm Scope The primary goal is the assessment of the reliability of the algorithm developed by [Bodnar et al., 2014]. The algorithm is based on a comparably small data set: 104 individual twitter users from the same university account for 37'599 tweets. Only 35 of these twitter users fell sick during the study period and tweeted a total of 1609 during the month in which they contracted influenza. Hence, we need to find out whether the algorithm can be used for

- different locations
- different time periods
- a more general Twitter population
- a bigger data set

Reliability Assessment Partially dependent on the results from the first goal. If applicability of algorithm could not be extended, we should of course assess the reasons for that. However, if scope of algorithm can be extended to other regions, time-periods and twitter populations, yielding reasonable results, we should assess some key statistical properties, in order to find out whether we can trust the results

- how stable is algorithm for extreme events?
- do algorithm parameters make sense from an epidemiological point of view?
- perform basic performance checks such as cross-validation on single models
- evaluate performance of meta-classifier
- how much do results from algorithm fit other epidemiological data, e.g. CDC ILI data
- compare algorithm with other web-based disease surveillance algorithms such as google flu trends (GFT)

Spatiotemporal Analysis of Dataset This last point is option. If - for which reason whatsoever - the achievement of the first two goals is not feasible or if we proceed faster than expected, we can use the data set in order to look at spatiotemporal differences in influenza-related tweeting behaviour and the propagation of influenza related tweets on the social network.

References

Todd Bodnar, Victoria C Barclay, Nilam Ram, Conrad S Tucker, and Marcel Salathé. On the ground validation of online diagnosis with twitter and medical records. pages 651–656, 2014.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

David D Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, pages 81–93, 1994.

Jason D Rennie, Lawrence Shih, Jaime Teevan, David R Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 3, pages 616–623. Washington DC), 2003.

Todd Bodnar and Marcel Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 699–702. International World Wide Web Conferences Steering Committee, 2013.

Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

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

Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, 107(41):17486–17490, 2010.

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

Declan Butler. When google got flu wrong. *Nature*, 494(7436):155, 2013.

Donald R Olson, Richard T Heffernan, Marc Paladini, Kevin Konty, Don Weiss, and Farzad Mostashari. Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in new york city. *PLoS Med*, 4(8):e247, 2007.