Potential of Social Media to Determine Hay Fever Seasons and Drug Efficacy¹

DE QUINCEY, Eda, PANTIN, Thomasb, KYRIACOU, Theocharsc and WILLIAMS, Nikkić

a School of Computing and Mathematical Sciences, University of Greenwich, London, UK, e-mail: e.de.quincey@gre.ac.uk
b Blackpool Teaching Hospitals NHS Foundation Trust, Blackpool, UK, e-mail: charles.pantin@bfwh.nhs.uk
c School of Computing and Mathematics, Keele University, Staffordshire, UK

Abstract – This working paper describes an exploratory study that has collected and analysed over 130,000 UK geolocated tweets from June 2012 to April 2013 that contained instances of the words “hayfever” and “hay fever”. Preliminary results indicate that the temporal and geographical distribution of tweets correlates with expected seasons and locations but allows for a finer level of granularity than currently available data sets. Common phrases that are being used and complaints relating to drug efficacy are also discussed as well as future planned analyses in areas related to risk management.

Keywords – Hay fever, Hayfever, allergic rhinitis, Social Media, twitter, eHealth.

1. Introduction

Hay fever or seasonal allergic rhinitis is a common allergic condition (Emberlin, 2010), defined as an Immunoglobulin E (IgE) mediated inflammatory response of the nasal lining following exposure to an allergen (Bousquet et al., 2008). Allergens include animal dander, house dust mite faeces, and grass and tree pollen with common symptoms comprising of nasal itching (pruritis), sneezing, nasal congestion and mucus discharge (rhinorrhoea). Other symptoms include itchy eyes (conjunctivitis), itchy throat (pharyngitis) and ears. The mainstay of treatment is avoidance of exposure to the causative allergen and symptom control with medications such as anti-histamines (van Cauwenberge et al., 2000).

The current UK hay fever prevalence is between 20-25% of the population, projected to rise to 39% by 2030 (Emberlin, 2010). The Royal College of General Practitioners (RCGP) Weekly Service Report Annual Report 2011 states that the mean weekly incidence of allergic rhinitis was 14.6 per 100,000 across all ages in 2011 (Fleming et al., 2011). Taking the 2011 census UK population estimate of 63.2 million, there were approximately 9.227 people with allergic rhinitis symptoms each week in the UK in 2011. Prescriptions for all nasal allergy rose from 2.7 million in 1991 to 4.5 million per year in 2004 (Gupta et al., 2007). Surges in incidence of allergic rhinitis in spring and summer are commonly known as the hay fever season, with the main pollens in the UK being birch pollen (March to mid-May) and grass pollen (late May to August) (Emberlin, 2010). However, determining an accurate start date of the season is difficult with Bielory predicting that the official pollen season in the U.S. will begin earlier in 2040 (April 8th) compared to 2000 (April 14th) (ACAAI, 2012).

Currently, the Meteorological Office (Met Office) provides weekly pollen forecasts and the Royal College of General Practitioners (RCGP) produce weekly service reports. However the former is predicative and the latter is dependent on sufferers reporting to their GP. For researchers and sufferers of hay fever, there is currently no method for identifying real-time, geolocated hay fever incidence.

A promising approach in the related field of Epidemiological Intelligence to detect seasonal illnesses is the use of Social Media (de Quincey & Kostkova, 2010). By collecting incidences of users self-reporting illnesses on twitter, it has been shown that outbreaks can be predicted 1-2 weeks before RCGP data indicates (Szomszor et al., 2012). The Kleenex™ tissue manufacturer Kimberly Clarke has used social media since 2011 to advise hay fever sufferers and promote its products. As part of this, they produced the “UK’s first real-time, interactive hayfever map” (Figaro Digital, n.d.) by encouraging people to ‘tweet’ the #hashtag ‘#atishoo’ followed by their postcode. Similarly, Anti-allergy drug Benadryl launched a social pollen count app “allowing users and other hay fever sufferers to report the

¹This article is based on a presentation given during the 2nd GRF Davos One Health Summit 2013, held 17-20 November 2013 in Davos, Switzerland (http://onehealth.grforum.org/home/)
pollen hotspots they encountered throughout their day” (Social Slurp, 2013).

However, these Social Media activities have often relied on users utilising specific, non-natural phrases within tweets and consequently have received little uptake and in the case of Benadryl have fallen victim to inappropriate posts and “graphic graffiti” (Social Slurp, 2013).

This working paper describes an exploratory study that has collected and analysed over 130,000, UK geolocated tweets from June 2012 to April 2013, that contained instances of the words “hayfever” and “hay fever”.

2. Methodology

Twitter allows access to users’ tweets via a number of Application Programming Interfaces (APIs). Following the same methodology previously used by one of the authors (de Quincey and Kostkova, 2010) to investigate the potential of twitter to predict Swine Flu outbreaks, twitter’s Search API was utilised. This allows developers to query tweets in real-time using a combination of keywords and other parameters such as a “geocode” which “returns tweets by users located within a given radius of the given latitude/longitude” (twitter, 2013).

The matching tweets, containing the text of the tweet, user information and a timestamp, are returned in either atom (an xml format) or json (a computer data interchange format). This data can then be parsed programmatically using a variety of computer languages such as PHP, Ruby, C etc..

2.1. Use of twitter in this study

For this study, the Search API was queried to return the last one hundred tweets that had a geographical location within the UK and contained instances of the words “hayfever” and “hay fever” (to allow for misspellings). To restrict returned tweets to those within the UK, the geocode parameter was set as a radius of 350 miles from the centre of UK and Ireland (“54.388,-4.536,350mi”).

It should be noted at this point that including the geocode parameter means that the tweets returned from the Search API are only those that either contain a specific longitude and latitude (which can be included within a tweet by the user e.g. via posting using their phone with geotagging enabled) or where the user has included a location within their Twitter profile. This means that tweets that mention “hayfever” or “hay fever” that do not contain a geographical location have not been collected and also that the location of some tweets might not be an accurate representation of where the user actually is when they posted the tweet e.g. a user has set their location as “London” on their profile but has sent the tweet whilst visiting Manchester.

PHP code was written to parse the returned tweets (in JSON format) and save unique records to a MYSQL database, comprised of one table. Records collected comprised of the following fields:

- id, created_at, from_user, from_user_id, text, source, geo, location, iso_language_code, profile_image_url, to_user_id, in_reply_to_status_id_str, query

A scheduled task was created that ran the PHP code every minute with new tweets being saved in the database. The program was started at 12:55 on 20th June 2012 and has being running continuously since then. The results presented in this extended abstract are taken from the starting time up to 08:47 on 2nd April 2013.

3. Preliminary Results

During the 286 day period under investigation, 130,233 tweets have been collected from 88,747 distinct users. Interestingly, the majority of these tweets (83.5%) are comprised of the misspelled version of the term i.e. “Hayfever”. It was also noted that 23,144 (17.8%) of these tweets can be classified as retweets, “a re-posting of someone else’s Tweet” (twitter 2012) and have been removed from the sample analysed for this study, as they are akin to duplicate entries.

3.1. Use of twitter in this study

Figure 1 on page 295 shows the distribution of tweets during the 286 day period.

The highest number of tweets posted was 5,826 on the 26th of June 2012 and the lowest was 4 on 31st March 2013. It is clear that there are peaks during this time period on certain days in June, July and August and then smaller peaks starting in March. The highest number of tweets from a single user was 413 (@WindowScreensUK who “supply fly, pollen, pet and solar window screens”). However, from all tweets, including retweets, the majority of users 67,341 (75.9%) only posted one tweet.

The distribution of these tweets for the time period analysed is similar to Pollen Calendars produced by the National Pollen and Aerobiology Research Unit1, with peaks in June/July, reductions through August/September, no pollen from October to January and then a rise in March.

3.2. Content of tweets

From all tweets collected, 29,955 (23%) contained hashtags. The most popular hashtags used included #hayfever (6,991); #itchyeyes (326); #fuckoff (293); #dy-ing (266) and #fml (264). As well as #itchyeyes, other hashtags relating to symptoms were being utilised such as #sneeze; #cantstop sneezing; #sneezing; #achoo; #soreeyes; #puffyeyes; #sneez; #sniff and #sniffles.

Frequently used self-reporting phrases found included “I have hayfever” (1,006); “I have hay fever” (332); “my hayfever” (6,707) and “my hay fever” (1,124). There were also a number of incidences of users self-reporting symptoms including phrases such as “my eyes” (3,068); “my

1Due to changes in twitter’s API, this code has now been expanded to include additional fields.
nose” (1,159) and “my throat” (203), followed by symp-
tom terms such as “watering”; “streaming”; “burning”;  
“stinging”; “itching”; “itchy”; “closing”; “sore”; “running”; 
“blocked” and “killing”.

Interestingly, in contrast to a previous study carried 
out into Swine Flu by one of the authors (de incey and  
Kostkova, 2010), the phrases used to self-report illnesses  
and symptoms are subtly different. When self-reporting 
flu, users tended to use a first-person subject pronoun  
such as “I”, or a third-person subject pronoun such as “he”  
or “she”, followed by the verb “have” or “has”. In this  
study, although users did use the phrase “I have hayfever”,  
more commonly they referred to “my hayfever”, using a  
possessive. Psychological ownership of illnesses has not  
been studied extensively in contemporary health and cul-
tural psychology (Karnilowicz, 2011) but it has been noted  
by Winkelman et al. (2005) that patients’ values and pri-
orities are often not reflected in medical records and sug-
gested that if patients keep a record of their illness expe-
riences from their own perspective this will enhance their  
sense of illness ownership. It could be argued that twitter  
and micro-blogging in general are now enabling sufferers  
to keep an informal log of their illness, increasing their  
own perception of ownership. This has advantages as it  
has been identified that patients who believe it is possible  
to control illness are more likely to adapt to the conse-
quences of the illness, attend rehabilitation programmes  
and adhere to treatment (Bridle, 2009). For example, in  
the case of a procedure such as a coronary bypass, per-
sonal control can be associated with a shorter hospital  
stay (Mahler and Kulik, 1990).

As well as self-reporting symptoms, 5,254 tweets re-
lated to medication were also found (containing terms  
such as “medicine”; “tablets”; “meds”; “medication”; “pill”;  
“spray” and “drugs”). The contents of these are currently  
being analysed further but a number contain general com-
plaints such as “tablets don’t work” and “the pills don’t  
work” (437). Preliminary textual analysis however seems  
to indicate that similar to the findings of Smith et al.  
(2012), some hayfever sufferers who have negative beliefs  
about their condition, believe they have little personal  
control over their illness and that their treatment is not ef-
efective. Sentiment Analysis is currently being considered  
as a method for determining whether negative sentiment  
in tweets relates to this perception and comparison with  
explicit methods of data collection such as the Revised Ill-
ness Perception Questionnaire (Smith et al., 2012).

3.3. Geographical distribution of tweets

The number of users who geotagged tweets was relatively  
low with only 3,924 tweets (3%) having a precise longitude  
and latitude. However, all tweets contained an approxi-
mate location e.g. 16,365 were posted from a profile that  
had a location set as “London”. The figure below shows the  
geographical distribution of all tweets collected within the  
UK and Ireland, made with the Google Maps Geoencoding  
service and Google Fusion Tables.
Figure 2: Distribution of geolocated tweets posted June 2012 to April 2013 containing the terms "hayfever" or "hay fever".

Analysis of this data is ongoing but visual comparisons with a map of UK Pollen Hotzones produced by the Met Office shows a similar distribution, with South England, the Midlands and South Wales showing high/very high prevalence of hay fever related tweets and pollen, with areas in the North East and North West of England and a band across the South of Scotland having moderate levels.

4. Future work

A number of areas for future work have been identified in previous sections but the following sets of analyses are also currently being initiated:

1. The mapped data will be normalised to the population living in the region collected i.e. the UK will be partitioned into a grid and then the number of tweets collected within those regions will be divided by the population in that square. Plotting these ratios on a map will provide allergen hotspots and reduce the effect of population density. This method can subsequently identify problem areas that can perhaps help identify concentration of problematic plants, animals, pollution levels (due to industrial processes etc.).

2. Following the same principle outlined above, the current dataset will be normalised to other geographical data sources such as vegetation maps, specifically the concentration of a specific plants known to cause hay fever.

3. Confidence intervals will be calculated to take into account the fact that twitter users are not necessarily representative of an entire population.

4. The 3 previous sets of analyses will be combined to produce a generic methodology for analysing geolocated data collected from twitter and applied to other areas of interest related to risk management of other illnesses and diseases.

5. Conclusions

The initial results of this study suggest, similar to previous experiments that have data mined twitter for health related information, Social Media can be used to determine people self-reporting illnesses and their symptoms implicitly (as opposed to explicitly asking people to self-report using specific phrases or hashtags). Although for hay fever, data sources are already available in the form of pollen counts from the Met Office and incidence rates from RCGP, these are either predicative or retrospectively aggregated. For sufferers, researchers and practitioners reliant on knowing when hay fever seasons occur e.g. ns, determining an accurate start date is difficult. The higher level of granularity that twitter enables for temporal analysis means that daily peaks of hay fever incidence can be identified in real time along with potentially pinpointing more accurate start and end dates of the seasons within different parts of the UK.

Although in this study the level of accurate geolocation data was low, geocoding general locations from user profiles has produced similar "hotzones" to those produced by the Met Office. In comparison, the data the Met Office provides to customers is daily and only for 18 locations within the UK (T Pantin 2013, pers. comm., 13 June).

This study has also shown that twitter enables researchers to access information regarding specific symptoms and medication usage/effectiveness. The terms and phrases identified have shown that sufferers are using hashtags and reporting common symptoms/complaints.

In conclusion, this study has shown that temporally accurate, geolocated hay fever incidence data is freely available via twitter. Although there are issues with the population sampled and the amount of noise present within the data collected, results similar to those available publicly regarding hay fever hotzones and seasons can be determined, along with information about symptoms and medication use.

References


van Cauwenberge, P; Bachert, C; Passalacqua, G; Bousquet, J; Canonica, GW; Durham SR et al. (2000): Consensus statement on the treatment of allergic rhinitis. European Academy of Allergology and Clinical Immunology. Allergy,55, 116-134.


Citation