#hayfever; A longitudinal study into hay fever related tweets in the UK

Ed de Quincey
School of Computing and Mathematics, Keele University
Keele, Staffordshire, UK
e.de.quincey@keele.ac.uk

Theocharis Kyriacou
School of Computing and Mathematics, Keele University
Keele, Staffordshire, UK
t.kyriacou@keele.ac.uk

Thomas Pantin
Macclesfield Hospital,
Macclesfield,
Cheshire, UK
tpantin@doctors.org.uk

ABSTRACT
This paper describes a longitudinal study that has collected and analysed over 512,000 UK geolocated tweets over 2 years from June 2012 that contained instances of the words "hayfever" and "hay fever". The results indicate that the temporal distribution of the tweets collected in 2014 correlates strongly ($r=0.97$, $p<0.01$) with incidents of hay fever reported by the Royal College of General Practitioners (RCGP) in the same year. An analysis of the content of the tweets indicates that users are self-reporting common, often severe symptoms as well as the uses of medication. We conclude that hay fever related tweets provide a real-time, free and easily accessible source of data at a finer level of granularity than currently available data sets. The implications for researchers, health professionals and sufferers are also discussed.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *Data Mining.*

General Terms
Measurement, Experimentation, Verification.

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

1. INTRODUCTION
Hay fever or seasonal allergic rhinitis is a common allergic condition [7], defined as an Immunoglobulin E (IgE) mediated inflammatory response of the nasal lining following exposure to an allergen [3]. 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 [22].

The current UK hay fever prevalence is between 20-25% of the population, projected to rise to 39% by 2030 [7]. 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 [9]. 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 [11]. 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) [7]. 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) [1].

Currently, the Meteorological Office (Met Office) provides weekly pollen forecasts and the RCGP produce weekly service reports. However the former is predicative and restricted to a limited number of specific locations and the latter is dependent on sufferers reporting to their GP. Both sets of data are publicly available but for a limited time with no freely available online archive of either type of data. For researchers and sufferers of hay fever, there is currently no method for identifying real-time (or accessing historical) geolocated hay fever incidence.

A promising approach in the related field of Epidemiological Intelligence to detect seasonal illnesses is the use of Social Media [4]. 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 [20].

The Kleenex™ tissue manufacturer Kimberley 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 (sic) map" [8] 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 pollen hotspots they encountered throughout their day" [19]. This service, supported by the Met Office, also now has a Mobile App, which combines "official Met Office data and live pollen alerts from sufferers to show exactly how pollen is behaving in your area" [13]. Users are asked to add "pollen hotspots" using the app, which automatically detects where the user currently is, and then rate how "bad" their symptoms are and what they think is causing the "problem". However, these Social Media based activities have often relied on users utilising specific, non-natural phrases within tweets or downloading specific applications. Consequently they have received little uptake and in the case of Benadryl have fallen victim to inappropriate posts and "graphic graffiti" [19] as well as technical issues [13] and user confusion.

Research by Gesualdo et al. [10], has shown that tweets containing allergic rhinoconjunctivitis symptom terms and the mention of an antihistamine drug show a high correlation with pollen counts from reporting stations in the US, over a 9-month period studied in 2013. They suggest that social media "may play a role in allergic disease surveillance and in signaling drug consumptions trends" [10].

Following on from an exploratory study that collected 130,000 hay fever related tweets during one hay fever season [5], this research paper describes the collection and detailed analysis of

over 512,000, UK geolocated hay fever related tweets from June 2012 to July 2014, covering three hay fever season cycles.

2. METHODOLOGY

Following the same methodology described in [5], twitter's search API has been utilised to collect the last 100 tweets from every minute, that had a geographical location within the UK and contained instances of the words "hay fever" and "hayfever" (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.

The program was started at 12:55 on 20th June 2012 and ran until the 31st July 2014[1]. The results presented in this paper are based on this 2-year period.

3. RESULTS

During the 772-day period under investigation, 512,198 tweets have been collected from 294,010 distinct users. Similarly to the smaller sample of 130,233 tweets analysed in [5], the majority of tweets collected, 69.4% (355,563), contained the misspelled vesion of the term i.e. "Hayfever". 108,468 (21%) can be classified as retweets "a re-posting of someone else's Tweet" [21] and have been removed from the sample analysed in section 3.1, as they can be considered as duplicate entries. 61,069 (12%) tweets were public messages to other twitter users, with the majority (99%) being in reply to a specific tweet.

3.1 Distribution of tweets

The distribution of all tweets is shown in figure 1 below, categorized into tweets and retweets.

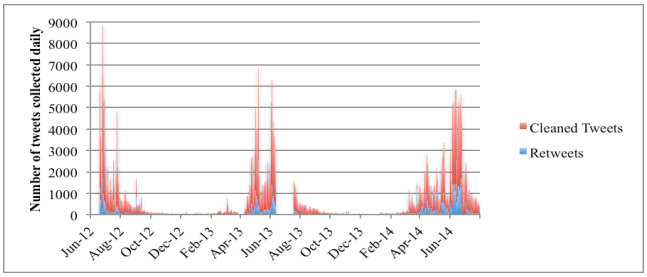

Figure 1: Distribution of geolocated tweets posted containing the terms "hayfever" or "hay fever".

The highest number of tweets posted was 5,826 on the 26th of June 2012. This figure however is skewed as 52% (3,002) of these were retweets of a tweet from a user, @carolineflack1, who currently has 1.7 million followers. A related phenomenon was seen at the end of March 2014 when a tweet by @GemmaAnneStyles, who currently has 3.16 million followers, was retweeted 1,353 times[2].

When discounting retweets, the distribution of tweets collected weekly during a calendar year is as follows. A moving average filter with a window size of 7 (one week) was applied to the data in order to remove weekday vs weekend effects.

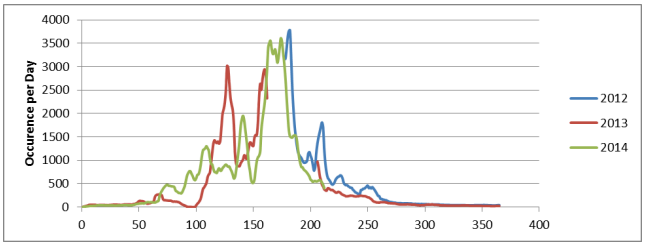

Figure 2: Distribution of geolocated tweets posted containing the terms "hayfever" or "hay fever".

It is clear that there are peaks during each year in June, July and August and then smaller peaks in March and April. The highest number of tweets from a single user was 9,875 (@OtrivineUK who sell nasal sprays and drops for the "symptomatic relief of nasal congestion"). However, from all tweets, including retweets, the majority of users 204,921 (70%) only posted one tweet.

3.1.1 Data Validation

In order to validate the data collected from twitter, a section (weeks 1 to 31 of 2014) was compared against the Allergic Rhinitis incidents reported in the RCGP weekly service report, available on the web for the same period [17][3]. Both time-series data were normalised to the interval [0,1] to allow for their comparison. Figure 3 shows how the two time-series compare.

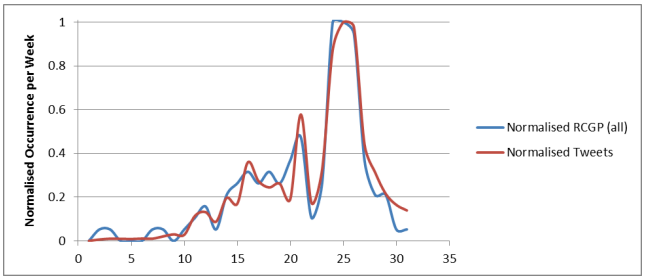

Figure 3: Comparison between the number of tweets with the RCGP reported data for the first 31 weeks of 2014.

The two time-series have a very strong correlation ($r=0.97$, $p<0.01$), which indicates that the hay fever related tweets have accurately reflected the changes in hay fever incidents reported by RCGP.

[1] Unfortunately due to hardware failures, results from 02/04/2013 to 09/04/2013 and 12/06/2013 to 18/07/2013 have not been recorded successfully and are not presented in this analysis.

[2] Both of these users have "affiliations" with the band One-Direction who have been ranked as the most influential users on Twitter in the UK [6].

[3] Data for the two previous years and for the rest of 2014 are unfortunately not publicly available via the RCGP. The authors have requested this data but no response was received from the RCGP by the time this paper was submitted.

3.2 Content of tweets

The most commonly used words found in all tweets related to hay fever such as "hayfever" (326,967 occurrences); "hay" (148,050) and "fever" (146,245). Other frequent words (discounting common words such as "a" and "is") included "my" (120,119); "like" (34,017); "today" (33,826); "eyes" (32,494); "hate" (30,964); "I'm" (28,436); "bad" (25,770); tablets (24,502); "fuck" (22,052); nose (19,859); "cold" (15,381); "summer" (15,248) and "sneezing" (14,429).

The following sections contain further analysis of the words and phrases used.

3.2.1 #Hashtags

Hashtags are unspaced phrases that are prefixed with the hash character (#) and are used on twitter to label tweets with groups and topics and allow for easier searching of related tweets.

From all tweets collected, 97,382 (19%) contained hashtags, the most common of which are shown in the following table.

Table 1: Most common Hashtags found in all tweets

Hashtag	Frequency	Hashtag	Frequency
#hayfever	30,447	#freebiefriday	1,096
#snotwhatisaid	10,282	#healthybrum	942
#itchyeyes	1,322	#allergies	923
#win	1,243	#dying	910
#pollen	1,123	#hayfeverproblems	886

In total, 26,268 different hashtags were used with #hayfever being the most popular. The hashtag #snotwhatisaid was predominantly used by one user, @OtrivineUK, who tagged all of their 9,875 posts with this hashtag. Interestingly, all of the tweets from this account were public replies to other users containing a link to their products (the equivalent of spamming).

Other popular hashtags often refer to symptoms, such as #itchyeyes (1,322); #sneezing (716); #sneeze (710); #achoo (550); #runnynose (427); #cantbreathe (400); #sneezy (342) and #puffyeyes (310).

3.2.2 Symptoms

According to [16], the symptoms of hay fever include frequent sneezing; runny or blocked nose; itchy, red or watery eyes; an itchy throat, mouth, nose and ears, and cough. In order to determine whether users are including symptom terms in their tweets, the occurrences of variations of the terms were counted e.g. for sneezing, the word stem "sneez" was searched for. The results of this are shown in the following table.

Table 2: Number of tweets containing symptom related terms

Symptom Word	Number of tweets
Eyes (itchy, red, watery etc.)	32,494 (6.3%)
Sneezing	26,391 (5.1%)
Nose (runny, blocked etc.)	19,859 (3.8%)
Itch (y,ing)	9,033 (1.7%)
Red	5,844 (1.1%)
Block (ed)	5,655 (1.1%)
Throat	5,025 (0.9%)
Water (y,ing,ed)	3,822 (0.7%)
Runn (y,ing)	3,449 (0.7%)
Cough	1,421 (0.3%)
Mouth	800 (0.2%)
Ears	574 (0.1%)

One of the other main symptom related terms found in 3.6% (18,368) of all tweets was variations on the term "kill" e.g. *"This hayfever is killin me man"*. Other symptom related words were also found such as "burning"; "stinging"; "closing" and "sore".

The words used around these symptoms, often demonstrate high levels of severity and discomfort for the poster e.g. *"Hayfever is totally kicking my arse today. The bastard."* Utilising a profanity dictionary [2], the number of tweets containing expletives was found to be around 11% (55,515).

3.2.3 Self-reporting and medication

In [20], it was found that self-reporting tweets containing phrases such as "I have flu", could be used as an early warning disease detection system. In this dataset, self-reporting phrases found included the following[4]:

Table 3: Number of tweets containing self-reporting phrases

Self-reporting phrase	Number of tweets
I have hayfever	4,804
I have hay fever	2,672
my hayfever	32,304
my hay fever	11,386
my eyes	14,103
my nose	5,413
my throat	1,835
my ears	138

The phrase "I'm" which also suggests self-reporting behaviour was user over 28,000 times e.g. "now I'm sneezing and crying"; "Feel like I'm dying". Further analysis of the phrases used around this word is currently ongoing.

The treatments for hay fever fall into 5 categories: Antihistamines; Corticosteroid nasal sprays and drops; Corticosteroid tablets; Nasal decongestants; Eye drops and Immunotherapy [15]. The term "Antihistamine" was present in 2,738 tweets, whereas "Immunotherapy" was only found in 31 tweets and "Corticosteroid" in 4. The following table shows occurrences of general terms for hay fever treatment mechanisms.

Table 4: Number of tweets containing treatment mechanisms

Treatment	Number of tweets
Tablets	24,502
Nasal/Nose spray	1,497
Eye drops/eyedrops	1,379

[4] A further discussion about why "My" is being used more than "I have" can be found in [13].

Nasal/Nose drops	125
Nasal decongestant	50

Interestingly, a number of users (around 670) when mentioning medication or treatments are commenting on their efficacy e.g. *"Why don't hayfever tablets work #irritating"*.

3.3 Geographical distribution of tweets

The number of users who geotagged tweets was relatively high with 38,480 tweets (7.5%) having a precise longitude and latitude. In [12], it is suggested that only 3% of all tweets contain native location information but the geo specific search criteria used for this study perhaps explain this higher percentage. All other tweets contained an approximate location set by the user within their profile e.g. 86,460 were posted from a profile that had a location that contained "London". In total, 27,064 distinct locations were recorded, but a number of these profile locations refer to the same place but use different terms e.g. "London, UK", "North London", "London, England" etc.

The following table shows the top 5 locations for number of tweets with similar location names combined (removing country names such as UK, England etc.), and the number of tweets divided by the population of each city according to the Office of National Statistics:

Table 5: Top 5 locations for number of tweets

Location	N^{o} of tweets	N^{o} of tweets/Population
London	86,460	1.03%
Manchester	21,489	4.18%
Birmingham	10,458	0.96%
Liverpool	10,206	2.17%
Bristol	8,244	1.88%

Further geolocation-based analysis is currently on going but initial analysis of the June 2012 to April 2013 data in [5], suggests that "visual comparisons with a map of UK Pollen Hotzones produced by the Met Office shows a similar distribution" to the geolocated twitter data.

4. DISCUSSION

4.1 Hay fever seasons

It is clear from the analysis presented in Section 3.1 that there is a very strong relationship between the hay fever related tweets collected and actual incidents of hay fever being reported by GP's within the UK for 2014. Cross correlation between the data collected during each of the three years (figure 2) shows identical peaks in June, July and August and smaller peaks in March and April. These findings are in line with the traditional hay fever seasons outlined in [14].

It is interesting that unlike in previous studies related to flu where the tweets occurred 1-2 weeks before incidents reported by the RCGP [4][20], the peak periods of hay fever incidents in this study closely match one another. This difference could be explained by the fact that hay fever is a condition whereas flu is a viral infection. Hay fever is a direct response to high levels of a particular allergen, which may prompt high numbers of sufferers to immediately report their symptoms (either at their GP's or via twitter). Flu spreads from person to person over a number of days/weeks/months and sufferers are actually advised to not go to their GP if they are otherwise fit and healthy. The fact that GPs may only see severe flu cases at the latter stages of their pathology, potentially in increasing numbers towards the latter stages of an outbreak may explain this temporal difference.

It is also worth noting that in [20], it was self-reporting tweets that were found to have a relationship with the corresponding RCGP data. Unlike [10], no content filtering has been undertaken in this study apart from the removal of retweets. It seems that the inherent nature of the search criteria i.e. the condition name and the location, filters out tweets that may create noise within the data set. One potential explanation may be that twitter accounts with a specific location are more likely to be of individual users as opposed to those of organisations and companies. When looking at the content of the tweets in Section 3.2, it is clear that the relevance of tweets to self-reporting though is reasonably high with common uses of words and phrases that imply personal suffering, symptoms and medication.

This close relationship between the twitter and RCGP hay fever data sets has significant implications for the usage of data from twitter as an additional source of information, not only for hay fever researchers but also for health professionals and sufferers.

4.2 Implications for researchers and health professionals

Previous attempts at crowd sourcing hay fever data from Social Media [8][19][13] have been unsuccessful, as users have been asked to provide data in specific formats or use a particular application to report their symptoms and location. This longitudinal study has shown that geolocated, real-time data is available via twitter that accurately reflects hay fever incidence and does not rely on these explicit data collection methods.

Currently, traditional hay fever data available for researchers and health professionals i.e. from the Met Office and RCGP, has limited granularity and accessibility. This study has also shown that twitter, with relatively simple search criteria, can be used as an accurate source of information for researchers and health professionals that has an added benefit of giving an indication of a more precise location. Although at this stage, we have not fully analysed the location data collected for all years, it is clear that there are opportunities to detect hay fever hotspots by taking into account population information. In Table 5 for example, it is worth noting that when dividing the number of tweets from a particular city by the population of that city, there are some cities that seem to have higher incidence than perhaps would be expected e.g. Manchester. As Section 3.2 has shown, there is also the potential for qualitative data analysis into the prevalence of symptoms and relative uses of medication and their efficacy.

A constant criticism of "big data" health studies related to twitter is the representativeness of the data, due to the inherent "youthfulness" of the twitter demographic [18]. For this study however, this effect was not seen, as both the distribution and relative volume of tweets, closely matched the RCGP data for all ages. This may be explained by the fact that sufferers "find that their symptoms improve as they get older" [16] and therefore self-reporting rates may decrease naturally with age as a consequence.

4.3 Implications for sufferers

For sufferers, although a hay fever calendar has been created by the Met Office [14], this does not give an indication of peak periods of hay fever incidence only when different types of pollen

that cause hay fever might be present (and a person may not know which type of pollen they are allergic to). The Met Office's daily pollen count report, does give an indication of what sufferers may experience that day but the location provided covers a wider area than is available via twitter. There is therefore a potential opportunity to combine this implicit data with the explicit data collected by [13] and pollen data to provide sufferers with a comprehensive update about current hay fever conditions.

5. CONCLUSION

Currently, future qualitative analysis into the content of the tweets themselves is planned along with the production of a comprehensive hay fever hotspot map, taking into account demographic data such as population statistics.

This study has shown though, that unique tweets containing the terms "hay fever" and "hayfever" posted by users in the UK, accurately reflect incidents of hay fever being reported by GP's. There are a number of opportunities for researchers, health professionals and sufferers in this area to use twitter as an additional, free and easily accessible source of data that has a number of advantages such as the finer level of temporal and geographical detail available.

6. REFERENCES

[1] American College of Allergy, Asthma and Immunology (ACAAI). 2012. The year 2040: Double the pollen, double the allergy suffering? ScienceDaily. Available at: www.sciencedaily.com/releases/2012/11/121109083736.htm [Accessed: 18 March 2015].

[2] BanBuilder. 2015. PHP profanity filter composer package for application developers, moderators, etc. BanBuilder. Available at: http://banbuilder.com/ [Accessed: 18 March 2015].

[3] Bousquet, J., Lund, V.J., Van Cauwenberge, P., Bremard-Oury, C., Mounedji, N., Stevens, M.T. et al. 2003. Implementation of guidelines for seasonal allergic rhinitis: a randomized controlled trial. Allergy. 58 (2003), 733-741.

[4] de Quincey, E., Kostkova, P. 2010. Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter. LNICST, Electronic Healthcare. Berlin Heidelberg: Springer. 27 (2010), 21-24.

[5] de Quincey, E., Kyriacou, T., Williams, N. and Pantin, T. 2014. Potential of Social Media to determine hay fever seasons and drug efficacy. Planet@Risk. 2(4), 293-297.

[6] Dredge, S. 2013. One Direction top chart of influential UK Twitter users. The Guardian. Available at: http://www.theguardian.com/technology/2013/oct/31/one-direction-top-chart-influential-uk-twitter-users [Accessed: 18 March 2015].

[7] Emberlin, J. 2010. The Hay Fever Health Report 2010. The National Pollen and Aerobiology Research Unit.

[8] Figaro Digital. n.d. Case Study: Kleenex. Figaro Digital. Available at: http://www.figarodigital.co.uk/case-study/Kleenex.aspx [Accessed: 18 March 2015].

[9] Fleming, D. M., Spofforth, N., Barley, M. A., Grant, S. J., Durnall, H. & Postle, H. 2011. Weekly Return Service Annual Report. Royal College of General Practitioners.

[10] Gesualdo, F. et al., 2015. Can Twitter Be a Source of Information on Allergy? Correlation of Pollen Counts with Tweets Reporting Symptoms of Allergic Rhinoconjunctivitis and Names of Antihistamine Drugs. T. Preis, ed. PLoS ONE, 10(7), p.e0133706.

[11] Gupta, R., Sheikh, A., Strachan, D. P. & Anderson, H. R. 2007. Time trends in allergic disorders in the UK. Thorax. 62 (2007), 91-6.

[12] Leetaru, K., S. Wang, A. Padmanabhan, and E. Shook. 2013. "Mapping the Global Twitter Heartbeat: The Geography of Twitter." First Monday. 18 (5).

[13] McNeil Healthcare UK Ltd. 2014. BENADRYL® Social Pollen Count. Available at: https://itunes.apple.com/gb/app/benadryl-social-pollen-count/id638068252?mt=8 [Accessed: 18 March 2015].

[14] Met Office. 2014. Pollen forecast. Met Office. Available at: http://www.metoffice.gov.uk/health/public/pollen-forecast [Accessed: 18 March 2015].

[15] NHS Choices. 2012. Hay fever - Treatment. NHS UK. Available at: http://www.nhs.uk/Conditions/Hay-fever/Pages/Treatment.aspx [Accessed: 18 March 2015].

[16] NHS Choices. 2014. Hay fever - Symptoms. NHS UK. Available at: http://www.nhs.uk/Conditions/Hay-fever/Pages/Symptoms.aspx [Accessed: 18 March 2015].

[17] RCGP Research & Surveillance Centre. 2014. NEW RSC Communicable and Respiratory Disease Report for England & Wales, Week 34. Royal College of General Practitioners.

[18] Sloan, L., Morgan, J., Burnap, P., Williams, M. 2015. Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLoS ONE. 10(3): e0115545

[19] Social Slurp. 2013. Benadryl Pollen Hotspot Map Goes Tits Up. Social Slurp. Available at: http://www.socialslurp.co.uk/benadryl-pollen-hotspot-goes-tits-up/ [Accessed: 18 March 2015].

[20] Szomszor, M., Kostkova, P., de Quincey, E. 2012. #swineflu: Twitter Predicts Swine Flu Outbreak in 2009. In: LNICST, Electronic Healthcare. Berlin Heidelberg: Springer. 69 (2012), 18-26.

[21] Twitter. 2013. GET Search. twitter. Available at: https://dev.twitter.com/docs/api/1/get/search [Accessed: 30 August 2013].

[22] van Cauwenberge, P., Bachert, C., Passalacqua, G., Bousquet, J., Canonica, G.W., Durham, S.R. et al. 2000. Consensus statement on the treatment of allergic rhinitis. European Academy of Allergology and Clinical Immunology. Allergy. 55 (2000), 116-134.