# Real-time monitoring of flu epidemics through linguistic and statistical analysis of Twitter messages

Karolos Talvis
Department of Informatics
Ionian University, Corfu, Greece
talviskarolos@gmail.com

Kostantinos Chorianopoulos
Department of Informatics
Ionian University, Corfu, Greece
konstantinos@gmail.com

Katia Lida Kermanidis
Department of Informatics
Ionian University, Corfu, Greece
kerman@ionio.gr

***Abstract* —The recent rise in popularity of Twitter and its open API provides developers the opportunity to extract amounts of data which can be a thesaurus of information. This opportunity led to the development of an open source and open API system called Flutrack (http://flutrack.org) that monitors influenza epidemics, based on geo-located self-reports on Twitter. In particular, we detect words such as *sore throat*, *cough*, *fever* etc. Moreover, we detect the aggravation of a patient's clinical condition when a user posts a second flu related tweet that contains words indicating further symptoms such as: *worse*, *deteriorating*. Finally, we present flu-positives with real time anonymous visualizations using maps (mapping), which might be helpful for authorities and sensitive populations to plan upcoming events or activities. In order to further aid the surveillance of the spreading of the disease, a classification experiment has been conducted for automatically identifying Tweets that describe cases with acute and more critical symptoms from those referring to milder cases. We found that making use of mereley very small n-gram keyword lexica, the automatic identification of critical cases reaches an accuracy of 92%.**

***Keywords—social media; Twitter; influenza epidemics; data visualization; data mining; linguistic analysis;***

## I. INTRODUCTION

Twitter, a micro-blogging service, has an estimated community of 500 million active users, generating 340 million tweets daily. Twitter users are enabled to send and read each other's 140-character messages, called tweets. Despite the high level of noise, the twitter stream does contain useful information, as it is useful for tracking or even forecasting trends, moods or behaviour if it can be extracted in an efficient manner. Furthermore, Twitter has been used as real time source for various public health applications. Our challenge was to create a real time application that tracks and visualizes influenza epidemics.

Influenza or flu is a viral infection that affects mainly throat, nose, bronchi and occasionally lungs. It is considered as one of the most common syndromes of infection in human beings. The symptoms are so common that self-diagnosis of Influenza is normal among the general public and clinical diagnosis. Flu differs from the common cold as it is caused by a different group of viruses and symptoms which tend to be more severe and last longer [1]. Infection usually lasts for about a week and is characterized by sudden onset of high fever, aching muscles, headache and severe malaise, non-productive cough, sore throat and rhinitis. Symptoms usually peak after two or three days. The best predictions for influenza are cough and fever, since this combination of symptoms has been shown to have a positive predictive value of around 80% in differentiating influenza from a population suffering from flu-like symptoms [1]. Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year [2][3]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza [4][5][6].

Our proposed system, Flutrack (http://flutrack.org), gathers flu related tweets in the English language for the entire world using the Twitter API [7]. The words used as tags are influenza synonyms and flu symptoms like *sore throat, cough, fever* etc. Further details are analyzed below. Another significant point is the localization of influenza epidemics aggravation. Since the above procedure is complete, we visualize and update tweets every 20 minutes. The Flutrack open-platform differs from similar tracking services as it extracts and processes without interruption flu related data in real time. This kind of characteristics makes the proposed service an ideal tool for extracting linguistic features and along with machine learning, for achieving accurate automatic recognition of patient's deteriorating symptoms and worsening conditions during flu infection [8][9].

Furthermore, making use of merely a set of very small n-gram, keyword and key-phrase lexica, and disregarding any kind of higher-level linguistic information (e.g. part-of-speech tags, syntactic structures etc.), tweets referring to cases with acute symptoms are distinguished from the ones describing milder cases, Such distinctions (also the identification of tweets referring to the author of the tweet as opposed

978-1-4799-6814-5/14 $31.00 © 2014 IEEE
DOI 10.1109/SMAP.2014.38

to tweets referring to a third party, or the identification of tweets referring to the disease in general rather than a specific patient in particular) are reported in the literature[11][23] and constitute significant support for monitoring the disease via social media. Unlike the reported approaches that usually rely on more sophisticated pre-processing of the tweet text, the approach proposed herein is more robust, and easily adaptable to other languages.

These characteristics make the proposed service an ideal tool for extracting linguistic features and along with machine learning, for achieving accurate automatic recognition of patient's deteriorating symptoms and worsening conditions, as well as during flu infection [8][9].

## II. RELATED WORK / RECENT APPROACHES

In 2010 Aron Culotta's proposed work investigated several models to analyze Twitter messages in order to predict rates of influenza like illnesses in a population [10]. He analyzed 500 million tweets from an eight month period and found that tracking a small number of flu related tweets, allows us to forecast future influenza rates with high accuracy, obtaining a 95% correlation with national health statistics.

In 2011 E.Aramaki, S.Maskawa and M.Morita addressed the issue of detecting influenza epidemics in large scale and in real time. The data were Twitter messages that included the simple word "influenza". The dataset was static, consisting of 300 million tweets extracted in a two-year period. A 5000 word training dataset was annotated by a classifier into positive or negative labels, depending on the tweet person or the surrounding persons that have the flu for positive label and the tense/modality tweets for negative label. Machine learning methods from the points of accuracy and time were used in order to lead to results [11].

Most recent approaches are devoted to developing applications that visualize data using mapping techniques or even foresee and anticipate influenza outspread. MappyHealth application searches Twitter for posts related with diseases, including influenza and visualizes them in a world wide scale. It also provides data which are open source [12]. In early 2013, Adam Sadilek visualized tweets in real time and analyzed how people living in polluted areas are more at risk of getting the flu than those living elsewhere [13].

It is essential for the Google Flu Trends' approach to be mentioned. The Google Flu Trends system monitors the flu activity of some countries and regions based on aggregated search queries from Google search engine. Estimates that came from that monitoring, have been validated through comparison with historic influenza data from the relevant country or region. In addition, Google Flu Trends uses IP address information from Google servers logs to make a best guess about where queries originated [14].

## III. PROPOSED APPROACH

Our challenge was to develop an open source and open data system that could extract Twitter messages related with influenza epidemics, filter and visualize them in real time. Except for the visualization, the flu tracking platform would be useful in extracting data destined to linguistic analysis that could result in an accurate prognosis of influenza outspread. The source code and the extracted/filtered data would be accessible to everyone in order to gain knowledge about flu detection or even transform and expand our project.

### *A. Flutrack platform*

Flutrack.org, the proposed system, examines Twitter data using the Twitter API. It gathers flu related tweets from the entire world, with searching tag, words that are influenza synonyms and flu symptoms. The detection of influenza is worldwide and only tweets in English language are being monitored. The tags that being tracked are: *Influenza*, *flu*, *chills*, *headache*, *sore throat*, *runny nose*, *sneezing*, *fever*, *dry cough*.

For every tweet extracted, additional amount of metadata is extracted too. Some of them are used for sorting the tweets in the database and other, such as geolocation coordinates, for mapping the tweets. Moreover, the Aggravation flag metadata define if a tweet is aggravated or not. For displaying tweets, Flutrack exports a JSON file from its database every 20 minutes. The exported file contains tweets from the last seven days. This process is showed below in (Fig. 1) and (Fig. 2) respectively.

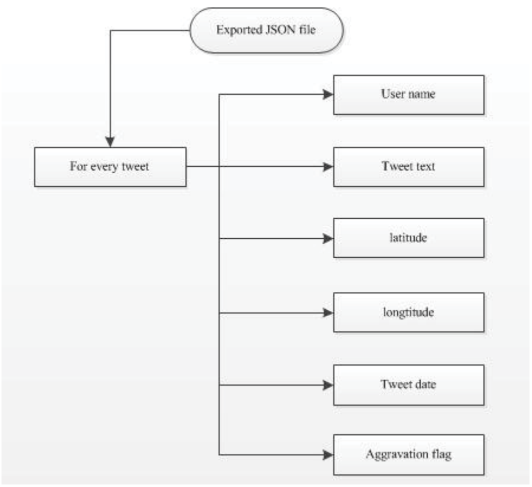

Fig. 1. Each tweet's metadata are shown above. Aggravation flag defines which marker will be used for mapping the current tweet, while the langtitude, longtitude are the coordinates in which the tweet is visualized. The Tweet date metadata specifies how long will the tweet visualized in the system.

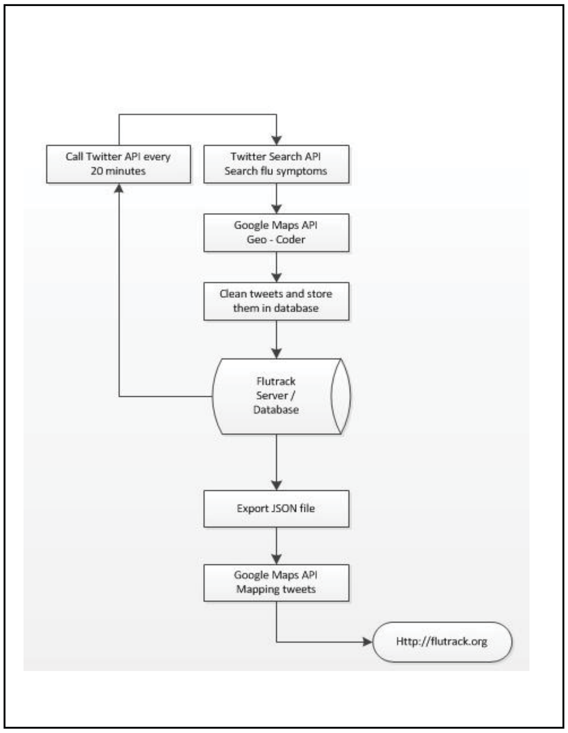
Call Twitter API every 20 minutes
Twitter Search API Search flu symptoms
Google Maps API Geo - Coder
Clean tweets and store them in database
Flutrack Server / Database
Export JSON file
Google Maps API Mapping tweets
Http://flutrack.org

Fig. 2. Overall flow of process.

In summary, Flutrack's platform architecture is based on speed and stability (Fig. 3). The PHP programming language is used for calling Twitter API, while JSON and Javascript for visualizing the tweets. The call requests are repeated periodically using a time based job scheduler named Cron. Finally, the whole application is hosted by the http://flutrack.org website using Jquery and HTML5 techniques.

### *B. Aggravation*

Aggravation of a patient's clinical condition is considered the case where a user posts a second flu related tweet that contains words indicating the aggravation of their condition such as: *worse*, deteriorating etc. In order to be registered as "aggravated", a tweet should be posted within seven days from the original. In case the same person re-experiences flu symptoms, for a period longer seven days, he/she is probably infected by a different virus, hence the tweet is listed as "common" and not "aggravated".

### *C. Linguistic filtering*

Before saving these tweets to the open database, the system filters them automatically by removing tags and hashtags (including @, # symbols). Moreover, only geolocated tweets and tweets that whose geolocation is extracted from the user's profile location (self declared home location) are being saved to the database. If enough location information is not available, tweets are automatically avoided. Tweets having less than 5 characters and those containing non-ASCII characters are excluded. Profile location is also automatically filtered from "suspicious" words (home, heaven etc), in consideration of avoiding false or non-existing location coordinates.

In addition to the linguistic filtering analysed above, a further filtering process is used to define which tweet post is referring to an influenza-infected user or not (flu positive/negative). Consider the following tweet examples:

- Flu positive: I'm definitely sick. Bad flu, cough, sore throat, fever & headache..
- Flu negative: Fever cough diarrhea upset stomach joint aches headache… I lost my appetite.

In the first case, the tweet contains the most common flu related symptoms caused probably by an influenza virus. As a result of this, this tweet will be displayed by the Flutrack platform as a common flu tweet.

In the second case, since the tweet contains the non-flu related symptoms "diarrhea" and "stomach aches", is tagged as flu negative. Although some kinds of flu may involve symptoms such as upset stomach or vomiting and diarrhea, the possibility of a different virus infection is very high. The infection could have been caused by food poisoning or gastroenteritis. Gastroenteritis is often called a "stomach flu" although it is not caused by influenza viruses [15][16]. In avoidance of displaying a non-flu related tweet to the map, the system automatically tags it as flu negative. However, there is always a small percentage of false positives tweets that are visualized, because of idioms, misspellings and shortened words the tweets contain.

## IV. API PROVIDER

An API (application programming interface) enables user to access a website's data without going near its databases. Flutrack's data are provided to every user via a simple JSON call. JSON (JavaScript Object Notation) is a lightweight, easy and popular way to exchange data. In order to facilitate users, Flutrack's API sorts and exports tweets according to:

- Flu symptoms
- Aggravation of patient's clinical condition
- Date and time
- Limit of tweets number

Every user of the API is able to synthesize a unique request to Flutrack's database based on his needs. The AND and OR operators are being supported for creating more complex queries. For example, if a user searches for tweets that contain the flu symptoms *fever* and *cough*, the corresponding request should be: http://api.flutrack.org/?s=feverANDcough. For the rest of the API choices, the same method should be used.

The whoe linguistic filtering procedure is developed with the PHP programming language by using multiple regular expressions, a sequence of characters that forms a search pattern in order to use pattern matching with string data (data which contain a finite sequence of symbols that are chosen from a set). Regular expressions are used automatically in Flutrack system in order to identify, filter or remove textual patterns that imported tweets contain as already analysed above.

## V. RESULTS AND STATISTICS

To evaluate the accuracy of our extracted data and results, a correlation between Flutrack's and Google Flu Trends' datasets had to be examined. All tweets, from both data providers, were geolocated in the U.S. and extracted from 12/2/2012 to 4/7/2013. The location was chosen based on the fact that the query counts of Google Flu Trends are compared with reliable sources, such as the U.S. Centers for Disease Control and other traditional flu surveillance systems for the U.S.

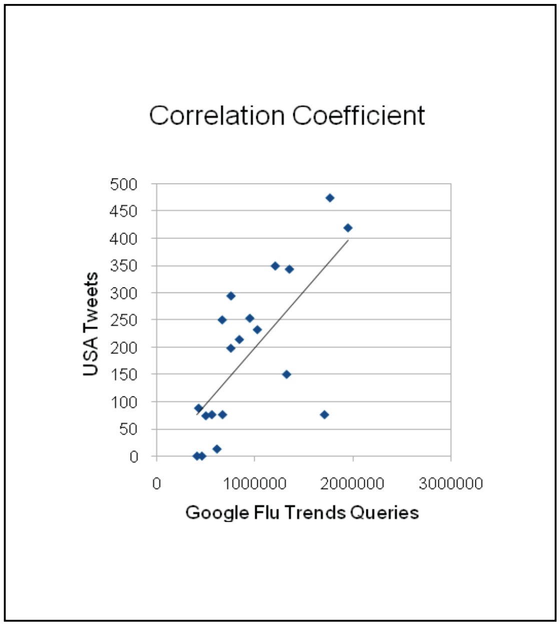

Fig. 3. Linear correlation coefficient between Flutrack and Google Flu Trends datasets.

In the figure (Fig. 3) above, the scatter plot of the population of tweets both from Google Flu Trends and the Flutrack platform is depicted. The straight line represents the correlation coefficient and the black dots the data of the two services. The correlation coefficient measure the strength of association between two variables. The most common correlation coefficient, called the Pearson correlation coefficient, measures the strength of the linear association between variables. The Pearson correlation has been used in our example. In order to quantify the degree of the observed linear relation between the uploads of the two social media, a linear regression analysis is also performed. In (Fig. 3) the least square line is depicted where the corresponding coefficient of determination is estimated to be 0.79 thus showing a high degree of correlation between Google Flu Trends and our platform.

## VI. IDENTIFICATION OF SYMPTOM INTENSITY

A classification experiment has been conducted for automatically identifying tweets that describe cases with acute and more critical symptoms from those referring to milder cases. Such a distinction helps improve the illness surveillance and monitoring process on social media [23]. The mining text has been proposed extensively in previous work for flu detection [15][16].

The classification schema has been based on three small manually crafted lexica that were developed by taking into account 700 tweets (training set). The first lexicon is a set of three word lists. When words from the second list are put together with words from the third list they form bi-grams (e.g. "incredible pain"). Words from the first list are optional and may precede the bi-gram, leading to tri-grams (e.g. "really incredible pain"). The second lexicon consists of multi-word expressions that denote intensity, such as "like hell", "like death" etc. The last lexicon consists of single words like "death" or "agony".

The remaining tweets (11448) were scanned so as to detect the appearance of a lexicon entry. The tweets that were found to contain a lexicon entry, i.e. that were classified as describing acute symptoms (test set), constituted 12.83% of the remaining tweets, and were manually evaluated as to whether they actually contained acute/critical symptoms. 92% were true positive cases, while the 8% false positives were cases where:

- either the entry referred to a third party (e.g. "**Fucked up** life it is." or "whoever got me sick should die"), and not to the situation of the sick person,
- or the context around the lexicon entry denotes possibility/tentativeness, e.g. "I seem to be **suffering**…" or "if I had a **really sore throat**…".

The matched lexicon entries are shown in bold. These results related to deeper tweet content analysis are quite promising, as they may influence disease surveillance, modeling author beliefs, disease awareness, as well as public health officials' response to outbreaks [23]. The aforementioned accuracy results are quite promising given the knowledge-poor nature of the classification approach and the extremely limited preprocessing tools required.

## VII. FUTURE WORK

In regard to statistical analysis, one step forward would be a study in time and space of the observed flue dynamics (evolution of infected population in different cities progressively in time) by means of recent multiscale models developed in other fields [17][18]. With these models, there is a possibility to use such social data as the ones available from the Flutrack platform, in order to achieve timely and robust detection of influenza epidemics.

One challenging (yet unexplored, to the authors' knowledge) research prospect would be the use of machine learning techniques to automatically detect the change in status of the disease, i.e. whether a patient's condition has deteriorated or improved, or to identify the critical symptom status. To this end, linguistic features (e.g. word unigrams or bi-gram phrases, comparative adjectives, quantifying adverbs etc.), emoticons and other social text features, like repeated characters (e.g. reaaaally), that reflect condition changes, and intensity could be employed for representing the status of a tweet.

Finally, making use of time and location stamps in tweets, the epidemics evolution can be modelled and its spreading predicted [18][19]. Related approaches have been proven to provide significant contribution to public health treatment.

## VIII. CONCLUSION

This paper proposed the detection of influenza outbreaks by processing and displaying flu related Twitter messages. For this exploration, an open source platform was developed. Our proposed system gathers and visualizes tweets every 20 minutes in real time. This open platform and its API allow users and developers to extend this project and take influenza detection to higher levels. The platform could work not only with Twitter but with any data provider. The proposed infrastructure might be employed by officials or individuals for planning their activities according to heat-maps of the flu epidemics. Moreover, in order to further help the monitoring of the disease, critical and intense symptom cases are identified from milder ones with high accuracy via tweet content analysis, using merely three small key-word and key-phrase lexica in an automatic classification experiment.

One step further would be a deeper statistical analysis to the correlated results that came from Flutrack, Google Flu Trends, and Centre for Disease Control and Prevention (CDC). Other parameters, such as weather deterioration or sudden warming statistics should also be analyzed, in order to examine if there is any correlation between weather changes and flu outbreaks. Moreover, by applying linguistic analysis and machine learning to Flutrack's results, it will be possible to make an accurate prognosis for influenza epidemics prevention and outbreak detection [20][21].

## REFERENCES

[1] Ronn Eccles, “Understanding the symptoms of the common cold and influenza ,” The Lancet Infectious Diseases, Vol 5, Issue 11, pp. 718-725, 2005.

[2] Karin Thursky, “Working towards a simple case definition for influenza surveillance ,” Jurnal of Clinical Virology, Vol. 27, Issue 2, pp.170-179, 2003.

[3] Ghendon, Y., “Introduction to pandemic influenza through history ,” European Journal of Epidemiology, pp.451-453, 1994

[4] Taubenburger JK, Morens DM, “Influenza: the once and future pandemic ,” Public Health Reports (Washington, D.C.:, pp.3:16-26, 2010.

[5] Beveridge W, “The Chronicle of influenza epidemics,” History and Philosophy of the Life Sciences, pp.13(2) : 223-234, 1991.

[6] Belshe RB, “An introduction to influenza: lessons from the past in epidemiology, prevention and treatment ,” Managed Care (Langhorne, Pa.), pp. 2-7, 2008.

[7] Twitter Developers Documentation, https://dev.twitter.com, 2013.

[8] Alexander Pak, Patrick Paroubek, “Twitter as a corpus for Sentiment Analysis and Opinion Mining” Proceedings of LREC, 2010.

[9] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Rassonneau, R., “Sentiment analysis of Twitter data,” LSM '11 Proceedings of the Workshop on Languages in Social Media, pp. 30-38, 2011.

[10] Aramaki, E., Maskawa, S., Morita, M., “Twitter Catches The Flu, Detecting Influenza Epidemics using Twitter ,” Conference on Empirical Methods in Natural Language Processing, EMNLP, 2011.

[11] Lamp A., Michael J. Paul, Dredze Mark, "Separating Fact from Fear: Tracking Flu Infections on Twitter". Proceedings of NAACL-HLT 2013, pages 789–795, Atlanta, Georgia, 9–14 June 2013.

[12] Aron Culotta, “Towards detecting influenza epidemics by analyzing Twitter messages ,” Proceedings of the First Workshop on Social Media Analytics, 2010.

[13] Brian Norris, Charles Boicey, Mark Silverberg, “MappyHealth application ,” http://mappyhealth.com , 2012.

[14] Adam Sadilek, Henry Kautz, “Modelling the impact of lifestyle on health at scale ,” Proceedings of the sixth ACM international conference on Web search and data mining, 2013.

[15] Google Flu Trends, http://www.google.org/flutrends/, 2013.

[16] Google Maps API V3, https://developers.google.com/maps, 2013.

[17] Courtney D. Corley, Diane J. Cook, Armin R. Mikler, and Karan P. Singh, “Text and Structural Data Mining of Influenza Mentions in Web and Social Media ,” International Journal of Environmental Research and Public Health, Vol 7, pp. 596-615, 2010.

[18] Collier, N., “Uncovering text mining: A survey of current work on web-based epidemic intelligence ,” Journal on Global Public Health, 2012.

[19] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu and Benyuan Liu, “Twitter Improves Seasonal Influenza Prediction,” HEALTHINF, 2012.

[20] M.. Avlonitis, E. Magkos, M. Stefanidakis and V. Chrissikopoulos, “A spatial stohastic model for worm propagation: scale effects”, J.Comput. Virol. 3:87-92, 2007.

[21] Adam Sadilek, Henry A. Kautz, Vincent Silenzio, “Predicting Disease Transmission from Geo-Tagged Micro-Blog Data ,” AAAI, 2012.

[22] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., “Detecting influenza epidemics using search engine query data,” Nature 457, pp. 1012–1014, 2009.

[23] Alessio Signorini, Alberto Maria Segre, Philip M. Polgreen, “The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic”, PLoS ONE, www.plosone.org , 2011.