*MEDINFO 2019: Health and Wellbeing e-Networks for All*
*L. Ohno-Machado and B. Séroussi (Eds.)*


*doi:10.3233/SHTI190366*

# Mining Social Media for Perceptions and Trends on HIV Pre-Exposure Prophylaxis

**Christine Kakalou[a], Jeffrey V Lazarus[b], Vassilis Koutkias[a]**

[a] *Institute of Applied Biosciences, Centre for Research & Technology Hellas, Thermi, Thessaloniki, Greece,*
[b] *Barcelona Institute for Global Health (ISGlobal), Hospital Clínic, University of Barcelona, Spain*

**Abstract**

*Pre-Exposure Prophylaxis (PrEP) is an approach for preventing the human immunodeficiency virus (HIV), which entails the administration of antiretroviral medication to high-risk seronegative persons. If taken correctly, PrEP can reduce HIV infection risk by more than 90%. The aim of this study was to identify and examine PrEP-related perceptions and trends discussed on Twitter. Using open-source technologies, text-mining and interactive visualisation techniques, a comprehensive data gathering and analytics Web-based platform was developed to facilitate the study objectives. Our results demonstrate that monitoring of PrEP-related discussions on Twitter can be detected over time and valuable insights can be obtained concerning issues of PrEP awareness, expressed opinions, perceived barriers and key discussion points on its adoption. The proposed platform could support public-health professionals and policy makers in PrEP monitoring, facilitating informed decision making and strategy planning for efficient HIV combination prevention.*

***Keywords:***
HIV; Data Mining; Social Media.

## Introduction

Pre-exposure prophylaxis (PrEP) is the administration of antiretroviral therapy to prevent seronegative persons from getting infected with human immunodeficiency virus (HIV). It is most commonly taken orally as a pill either daily or on demand, but other forms are also being tested in clinical trials (e.g. vaginal gels, injectables, and implants). The World Health Organization (WHO) recommendations state that individuals at high-risk should be offered PrEP as an additional prevention choice, as part of a broad HIV prevention strategy [1], outlining five key populations at-risk for contracting HIV [2], i.e. (a) men who have sex with men (MSM), (b) people in prisons and other closed settings, (c) people who inject drugs, (d) sex workers, and (e) transgender persons.

The rapid spread of social media has changed the way that people share their personal health experiences with others as well as the manner of seeking disease information and treatment alternatives [3]. Through social media, huge volumes of data are constantly being generated, providing a significant data source for research. In particular, patient-generated data referring to personal experiences in a way that is unsolicited, spontaneous and up-to-date, in large volumes and with a high degree of variance, is of great interest for research on public health surveillance [3]. Social networks focusing on health-related topics (e.g. medication and treatment options, side-effects, etc.) are particularly useful, since they are providing a wealth of information that includes strong semantic associations. The exploitation of social media facilitates the need to "hear the patient's voice", providing a unique opportunity to obtain input that cannot be acquired through official healthcare channels [4], either because certain groups don't have access to them, or because people are reluctant to provide sensitive information.

However, mining social media presents considerable challenges, which are linked with the inherent characteristics of user-generated data, e.g. vastness and no structure, recency, uniqueness, frequency, and salience [5]. The analysis techniques, relying typically on text mining/Natural Language Processing (NLP), need to be adjusted to the intricacies of the short, informal and colloquial text that people use to express themselves when posting online [6]. Specifically, in health-related use cases, the use of layman, descriptive, subjective and often misspelled terms complicate NLP tasks such as syntactical analysis, tokenizing and direct matching with medical lexicons. Supervised learning algorithms are rendered more difficult by the high amount of noise in the frequently vast collection of data, which demands more computational power and poses significant problems concerning performance.

Twitter is a very popular microblogging and social networking platform, through which users can post "tweets" i.e. short messages consisting of 280 characters maximum. Twitter is a common data source for opinion and knowledge mining on public health topics [5], such as pharmacovigilance [7], [4], prescription drug abuse [8], epidemiology surveillance [9], chronic diseases [10], mental health [11], and quality of care in hospitals [12] to name a few. One of the key barriers in sexual health communication is the patients' reluctance to seek information through their close personal networks [13]. Thus, the perceived anonymity offered by social media allows users to discuss sexual health issues and share their experiences within their online social circles.

This work aimed to examine whether PrEP-related trends and useful information can be identified in Twitter discussions by using text-mining techniques. While Twitter data have been exploited for PrEP-related analysis [13]-[15], the current study elaborated on systematizing and reinforcing such analyses through a comprehensive Web-based platform integrating various text-mining and visual analytics methods. In particular, an Exploratory Data Analysis (EDA) was employed, in order to detect the main discussion topics, the most frequent terms and popular hashtags along with possible side-effects regarding PrEP. In addition, sentiment analysis was used to investigate the public awareness on PrEP, the views and emotions towards it, as well as potential barriers for its adoption. The geographic distribution of the collected tweets along with the sentiment analysis scores were assessed, in order to discover potential insights on regions with low PrEP awareness or where PrEP meets obstacles for its acceptance.

In the following, we present the employed methods for the development of the platform, the obtained results, and discuss our findings along with future work directions.

## Methods

### Data Acquisition

Data acquisition relied on the SEST platform, an in-house developed software application [7]. SEST focuses on the systematic data aggregation and exploitation from various sources of unstructured text through NLP techniques, aiming to support public health surveillance applications. Built upon the microservices architecture, it enables the repurposing of its modules to accommodate alternative workflows/scenarios, interoperability with other systems through standards-based data exchange and semantic annotation of NLP results for further analysis.

Using SEST, which exploits the Twitter Streaming API [16], tweets were collected from September 2017 to December 2017 through two different collection streams:

1. "PrEP Datastream": tweets gathered with keywords: (truvada) OR (hiv AND PrEP) OR (preexposure AND prophylaxis AND hiv)
2. "Pill Datastream": tweets gathered with keywords: (pill AND prevent AND hiv) OR (pill AND protect AND hiv) OR (pill AND protect AND AIDS)

The "Pill Datastream" was set up to obtain tweets related to PrEP from users that might be unaware of the terminology of PrEP and/or its commercial name, i.e. truvada. The collected data were stored in MongoDB [17], an open-source, flexible and scalable NoSQL database that stores data in JSON format, while the RESTHeart API [18] was used for the create, read, update and delete (CRUD) operations in the SEST platform. For each gathered tweet, its content was stored in JSON format along with its unique id, the creator name, and the creation timestamp. For geolocation, only the location stated in the creator's bio was stored and not the geographical coordinates for each tweet (if available), aiming to preserve the creator's privacy. In addition, only tweets written in English were collected.

### Data Analytics

Various types of analytics were applied in the collected tweets based on text-mining and NLP techniques implemented in R (Figure 1). In the preprocessing stage, punctuation and stopwords were removed from the text and special points of interest, namely, hashtags (i.e. #PrEPworks), mentions (i.e. @UNAIDS) and URLs, were detected using regular expressions. Through the Term Frequency - Indirect Document Frequency (TF-IDF) normalisation algorithm, term frequencies of the corpus were calculated and word associations with a correlation threshold of 0.2 were presented. The main discussion topics and the 7 most prevalent terms per topic were detected using the Latent Dirichlet Allocation (LDA).

Sentiment analysis was performed to discover the polarity of expressed opinions in the corpus overall, but also per tweet and per creator. For the polarity calculation, the AFINN lexicon was used that has been especially developed for sentiment mining of microblogging text, featuring also slang words [19]. The emotions expressed within the corpus were mapped to the 8 basic human emotions (namely, anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) via the NRC Word-Emotion Association Lexicon [20]. Both lexicons had to be adjusted to exclude terms that skewed the polarity, such as "negative" and "positive", which in this specific context refer to the HIV status of an individual, and to include colloquial terms and abbreviations often mentioned in the collected tweets (i.e. "ftw", which means "for the win", and "dope" meaning "great").

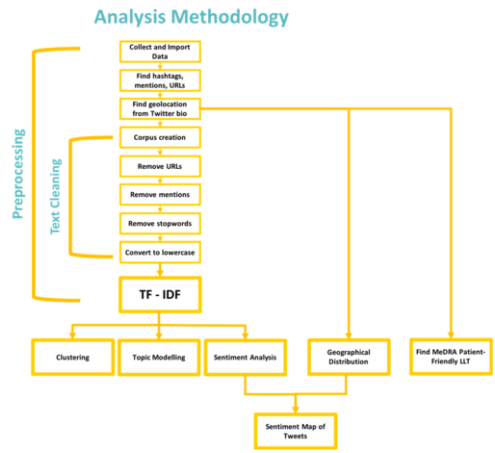


*Figure 1 - Analysis Methodology*

Since more than 23,000 tweets contained location information in the users' bios, they were extracted and converted to coordinates via the Google Geolocation API [21]. Locations that returned more than one set of coordinates or that did not correspond to real locations were rejected.

Possible mentions of side effects were identified by matching terms in the corpus of the recently released MedDRA® patient-friendly Low-Level Term (LLT) lexicon [22], which contains 1,440 terms that patients use to describe Adverse Drug Reactions (ADRs) and maps them to MedDRA® terms.

EDA was then applied to enhance the understanding of the collected data, uncover underlying trends, outline important variables, and detect outliers/unexpected behaviours that might signify important insights regarding the research questions. Visualisation plays a pivotal role in EDA, as it enables intuitive pattern recognition, when studying public health trends over time.

The Shiny package in R was used for the visualisation of the analysis; an interactive Web application (Figure 2) was deployed online, providing an overview of the acquired data along with suitable visualisation techniques for each aspect of the analysis and the ability to select text-mining parameters such as frequency and correlation thresholds where applicable.


*Figure 2 - Homepage of the PrEP trends monitoring platform*

## Results

Over the course of 4 months, we identified more than 32,000 PrEP-related tweets (31,854 from the "PrEP Datastream" and 804 from the "Pill Datastream"), which constitute only a small fraction of the total Twitter datastream at a given time. 60.5% and 69.3% of the tweets were retweets in the "PrEP Datastream" and the "Pill Datastream", respectively. We expected that the "Pill Datastream" would contain original content from people seeking information about PrEP, but the bulk of

tweets came from institutional or NGO campaigns that propagated messages like "PrEP is a pill that you can take once a day to prevent HIV".

In the following, we present the results of each analysis step.

**Frequent and Associated Terms**

Using TF-IDF normalisation the most frequent terms address prevention, the key at-risk populations (i.e. "msm", "gay", "trans"), and countries or cities where PrEP becomes available such as Canada, Ireland, and the United Kingdom. Generic PrEP as well as free access to it or insurance coverage was similarly repeatedly discussed. Also notable was the high number of tweets discussing clinical trials and, more specifically, the inclusion of a more diverse sample, such as transgender people and women from different countries. The PrEP discourse within Black African communities seems to diverge, ranging from strong acceptance as a tool of empowerment among black women to a high level of doubt and stigmatisation within the black MSM community.

Word associations display the most similar terms to a selected word of interest above a user-defined threshold. The terms more closely associated with "buy" reveal that people seek to buy truvada "otc" meaning "over the counter" and find the "cheapest" options such as "viread", another antiretroviral drug.

**Topic Modelling**

Using the LDA algorithm, we identified the 50 and 30 top discussion topics for the "PrEP Datastream" and the "Pill Datastream", respectively, with 7 terms per topic shown. For the visualisation, the LDAvis [23] method was used, which presents the topics globally and reveals how much they differ from each other. All topics discussed were found relevant to PrEP and tend to cluster together in large numbers, which is expected considering the very similar context in all PrEP-related tweets. Smaller clusters and topics that are more separate from the large cluster represent specific events/tweets that were largely propagated in the Twitter stream, such as topic #21 (Figure 3) that contains the terms "thanks", "uk", "cases", "percent", "men", "plunge", "gay", and clearly refers to the impressive decline in HIV infections amongst MSM in London after PrEP was made available in clinics.

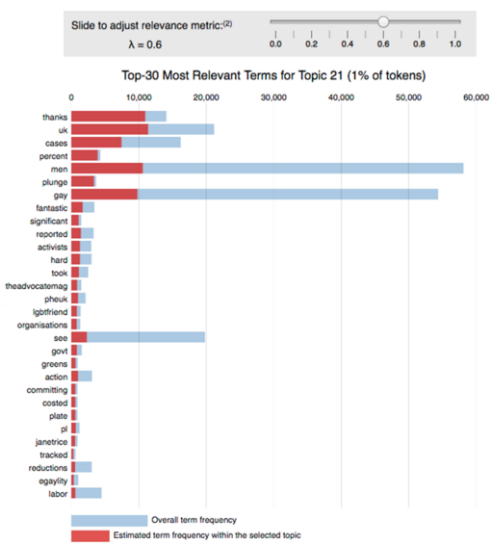


*Figure 3 - Most relevant terms for topic #21 using the LDA algorithm*

**Sentiment Analysis**

The overall sentiment was found positive, scoring 0.208 and 0.328 in a scale of -1 (absolutely negative) to +1 (absolutely positive) for the "PrEP" and "Pill" datastreams, respectively. The sentiment score densities calculated for each datastream show a more uniform distribution of sentiment for the "PrEP Datastream", while the "Pill Datastream" features a significant cluster of negative sentiment that, when traced back, reveals tweets expressing doubt and cautiousness regarding the existence of a pill that can prevent HIV infections. The most commonly expressed emotions in the corpus are trust and anticipation, with fear being very prominent in the "Pill Datastream".

**Hashtags**

Hashtags are a pivotal element of Twitter data analysis; not only do they signal emphasis in the context of a tweet text, they also act as nodes within the network, connecting various users and posts through a shared discussion topic. The use of a specific hashtag for a cause, event or awareness campaign is a common communication strategy adopted by organisations worldwide. The hashtags in the corpus were separately analysed and presented in a streamgraph visualisation. As expected, around the 1st of December (Word AIDS Day) a substantial peak was found for #WorldAIDSDay, #WAD2017, but also for HIV-related terms such as #TasP (Treatment as Prevention), #condoms, #EndTheEpidemic, etc. While this observation is evident, it serves as a validation control for accurately identifying trends over time using data-mining techniques. Observations of smaller peaks offer insights on events that influence the audience's perception and awareness of PrEP, such as the #Ireland hashtag that peaked after November 9th, 2017 when Ireland's first clinic monitoring PrEP was opened (Figure 4).

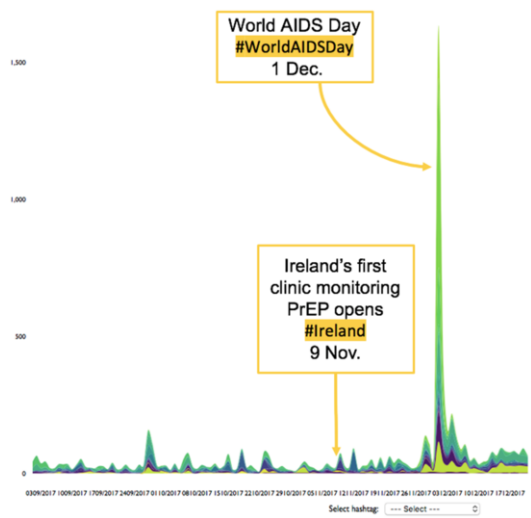


*Figure 4 - Hashtags timeline*

**References to Side-Effects**

The identified mentions to side effects that matched terms in the MedDRA® LLT lexicon concerned notably references to headaches, miscellaneous pain, and flu symptoms. A lot of noise was found in this mapping, because some lexicon terms are used in a different context than side-effect reporting, namely, "fatigue" for the phenomenon of "pill fatigue", "infection" as in "HIV infection", "worry" for being anxious about contracting HIV, and not as a result of taking PrEP medication.

It was evident that this type of analysis requires more sophisticated analysis and further research.

### Geographic Distribution

The geographic distribution of tweets, in the form of an interactive map, holds special interest, with the largest clusters of PrEP discourse being in the USA, Western Europe, and South Africa. Notably, the distribution shows a similarity with the UNAIDS 2016 map of countries [24] that have either approved PrEP for commercial use, or are in the pilot phase (Figure 5). There is an apparent lack of PrEP-related tweets in Russia, China and Japan that although probably attributed to the language barrier and the existence of alternative social networks to replace the - often blocked entirely - western social media, may also hint at clusters of populations with intense HIV stigmatisation, especially considering the high HIV prevalence and increased rate of infection in those countries.


*Figure 5 - Geographic distribution of tweets, and UNAIDS map of PrEP availability*

The sentiment score for each tweet was calculated and is displayed on an interactive "sentiment map" (Figure 6). Although significantly different clusters of positive or negative tweets are not evident, there is a slightly more positive concentration of tweets in the UK, Ireland, and the USA.

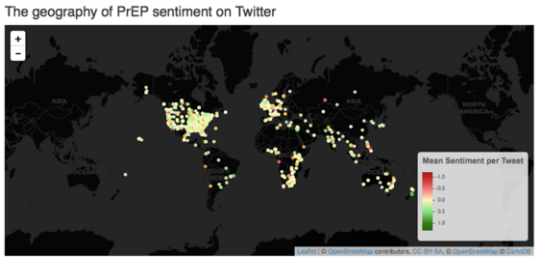


*Figure 6 - Geographic distribution with color-coded sentiment score per tweet*

### Influencers

Twenty-three (23) Twitter accounts per datastream, the vast majority corresponding to PrEP advocacy groups, LGBT networks and sexual health organisations, were found to post heavily about PrEP, representing more than 9% of the total posts collected. Those accounts constitute "PrEP Influencers", since they contribute to the dissemination of PrEP awareness across their networks and have a large number of followers (>10k mean follower count). The number of tweets per creator and per datastream as well as the sentiment score for the respective tweets is presented with a Treemap visualisation (Figure 7). Some accounts show negative sentiment associated with their posts as they mention issues like stigmatisation or feelings of guilt and isolation that may accompany a PrEP regimen.

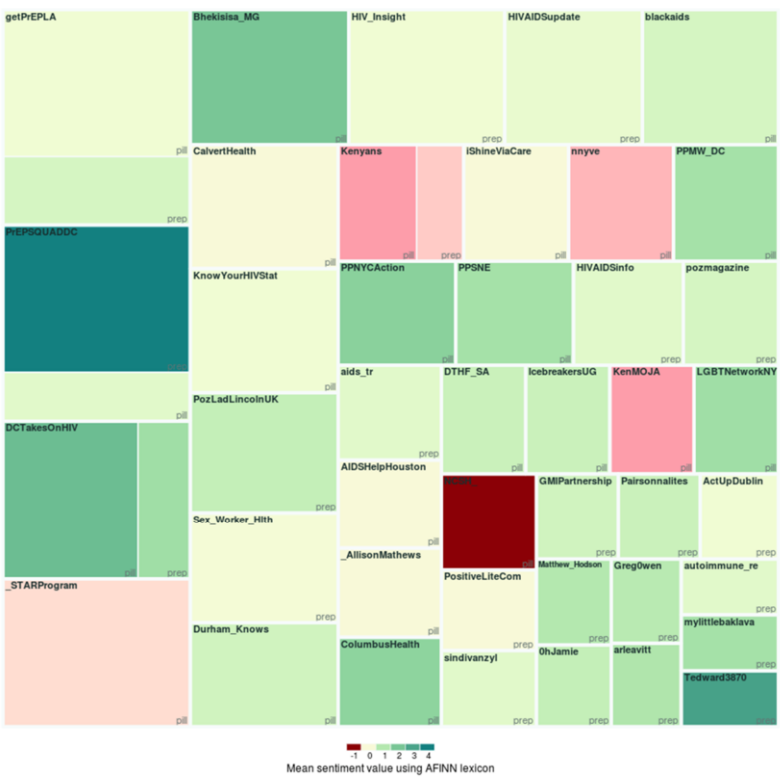


*Figure 7 - Treemap of top 23 influencers and their expressed sentiment score for each datastream*

## Discussion and Conclusion

The collection of tweets during the 4-month period proved to be quite prolific and with very low amounts of noise, indicating that PrEP discussions can be identified successfully on Twitter. PrEP discourse seems to be continuously increasing on Twitter and in a rather fast rate, considering the collected data over time. Users disseminate information about PrEP, engage in sharing ideas, experiences and considerations about it, while activists, HIV support and PrEP advocacy groups appear to drive the demand for free and accessible-to-all PrEP worldwide.

Overall sentiment regarding PrEP is positive and the main concerns refer to stigma, medication cost, insurance coverage as well as vulnerable population groups with limited access to PrEP. The various interactive visualisation techniques that we employed allowed for a more flexible inspection and exploration of the collected data. Detection of critical events that impact the target audience's reaction to PrEP is possible by combining the monitoring of trending hashtags, frequent terms and most discussed topics on the interactive platform provided.

The preliminary, shallow text-mining analysis provided satisfactory results in some context, such as the identification of the most frequent terms and word associations, as well as hashtags monitoring. However, other aspects of the analysis, e.g. side-effect mentions and the sentiment analysis would greatly benefit from more sophisticated NLP tasks, especially considering the unofficial/layman, brief, full of typos and abbreviations language contained in tweets. Besides addressing the above limitations, Machine Learning methods is another necessary future step to enable the accurate detection of personal vs informative tweets, a task which constitutes one of our future work directions. To effectively address this requirement, we are currently exploiting a significantly larger data corpus, which was acquired through our platform.

Overall, our results indicate that the current work provides the foundations for a comprehensive analytics platform that can offer insights to public health professionals and policy makers, so as to assist them in taking informed decisions and developing efficient strategies for improving HIV prevention.

## References


[1] WHO | Pre-exposure prophylaxis, *WHO*. (2018). https://www.who.int/hiv/topics/prep/en/ (accessed November 19, 2018).

[2] WHO, Cascade data use manual: to identify gaps in HIV and health services for programme improvement., Geneva, 2018.

[3] R. Harpaz, et al., Text mining for adverse drug events: the promise, challenges, and state of the art, *Drug Saf.* **37** (2014), 777-90.

[4] M. Salath, Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health, *J Infect Dis.* **214** (2016), S399-403.

[5] A. Sarker, et al., Utilizing social media data for pharmacovigilance: A review, *J Biomed Inform.* **54** (2015), 202-12.

[6] A. Sarker and G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, *J Biomed Inform.* **53** (2015), 196-207.

[7] P. Natsiavas, N. Maglaveras, and V. Koutkias, A public health surveillance platform exploiting free-text sources via natural language processing and linked data: application in adverse drug reaction signal detection using PubMed and Twitter, in: D. Riaño, R. Lenz, and M. Reichert (Eds.), *Knowl Represent Heal Care*, Springer International Publishing, Cham, 2017: pp. 51-67.

[8] C.L. Hanson, B. Cannon, S. Burton, and C. Giraud-Carrier, An exploration of social circles and prescription drug abuse through Twitter, *J Med Internet Res.* **15** (2013), e189.

[9] M. Santillana, A.T. Nguyen, M. Dredze, M.J. Paul, E.O. Nsoesie, and J.S. Brownstein, Combining search, social media, and traditional data sources to improve influenza surveillance, *PLoS Comput Biol.* **11** (2015), 1-15.

[10] M. Roccetti, et al., Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts, *JMIR Public Heal Surveil.* **3** (2017), e51.

[11] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, Multi-class machine classification of suicide-related communication on Twitter, *Online Soc Networks Media* **2** (2017), 32-44.

[12] J.B. Hawkins, et al., Measuring patient-perceived quality of care in US hospitals using Twitter, *BMJ Qual Saf.* **25** (2016), 404-13.

[13] M.L. McLaughlin, et al., Propagation of information about Pre-Exposure Prophylaxis (PrEP) for HIV prevention through Twitter, *Health Commun.* **31** (2016), 998-1007.

[14] P. Breen, J. Kelly, T. Heckman, and S. Quinn, Mining Pre-Exposure Prophylaxis trends in social media, in: *Proc IEEE Int Conf Data Sci Adv Anal*., 2016, pp. 214-21.

[15] J. Schwartz and J. Grimm, PrEP on Twitter: information, barriers, and stigma, *Health Commun.* **32** (2017), 509-16.

[16] Twitter Streaming API, (n.d.). https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/connecting.html.

[17] What Is MongoDB? | MongoDB, https://www.mongodb.com/what-is-mongodb (accessed February 18, 2018).

[18] Documentation - RESTHeart - Confluence, (2016). https://softinstigate.atlassian.net/wiki/spaces/RH/pages/9207957/Documentation (accessed March 10, 2018).

[19] AFINN, (n.d.). http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010 (accessed February 24, 2018).

[20] NRC Emotion Lexicon, http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed February 25, 2018).

[21] The Google Maps Geolocation API | Google Maps Geolocation API | Google Developers, (2018). https://developers.google.com/maps/documentation/geolocation/intro (accessed March 7, 2018).

[22] Patient-Friendly Term List | MedDRA, (2018). https://www.meddra.org/patient-friendly-term-list (accessed February 27, 2018).

[23] C. Sievert, and K.E. Shirley, LDAvis: A method for visualizing and interpreting topics, in: *Proc Work Interact Lang Learn Vis Interfaces*, 2014: pp. 63-70.

[24] UNAIDS, Fact sheet - Latest statistics on the status of the AIDS epidemic | UNAIDS, (n.d.). http://www.unaids.org/en/resources/fact-sheet (accessed May 5, 2017).



**Address for correspondence**

Christine Kakalou, Institute of Applied Biosciences, Centre for Research & Technology Hellas, 6th Km. Charilaou-Thermi Road, P.O. BOX 60361 GR – 57001, Thermi, Thessaloniki, Greece, email: ckakalou@certh.gr.