Using Twitter to study correlation between nutrition and health

Andreas Kalaitzakis

Thesis submitted in Partial Fulfillment of the Requirements for a

Master of Science degree (M.Sc.) in Information Systems

University of Crete
School of Sciences and Engineering
Department of Computer Science
Voutes Campus, 700 13 Heraklion, Crete,Greece

In cooperation with:

Laboratoire d' Informatique de Grenoble (LIG)

Thesis Advisor: *Assistant professor (MdC)* ***Noha Ibrahim***

Thesis Co-Supervisor: *Professor* ***Sihem Amer Yahia***

UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE

Using Twitter to study correlation between nutrition and health

Thesis submitted by

Andreas Kalaitzakis

in partial fulfillment of the requirements for a

Master of Science degree (M.Sc.) in Information Systems

THESIS APPROVAL

Author: _____

Andreas Kalaitzakis

Committee approvals: _____

_____

_____

_____

Grenoble, September 2014

Abstract

The last two decades we became witnesses of a rapid development of distributed computing and computer networks. Users that were initially restricted to access static text data that was available on the Internet are now enjoying multimedia content that is even produced by other users in real-time. The increasing proliferation and affordability of Internet devices, as well as the ease of publishing, searching and accessing information on the web encourages the individual users to communicate their content with the web society giving birth to the idea of social interaction imposing a growing need for systems that can extract useful information from this amount of data. One of the fundamental problems that emerged in social media stream analysis with a wide range of applications is to effectively detect underlying topics and their associated documents. It becomes clear that modern social services and social media show a substantial potential of providing society with a rather promising source of information which prevails over the traditional ones on a series of important dimensions. Recruiting social media in order to inform the public has proved to have a significantly lower operating cost in conjunction with a better propagation velocity. These advantages encouraged the academic community to investigate a framework under which a partial replacement of the traditional sentinel surveillance services with web enhanced ones could take place. Mobilized by this emerging need and recognizing a significant void in empirical studies that focus on nutrition, we randomly collected more than 200 millions tweets along with a series of accompanying features in a two months' period. Applying state of the art text analysis techniques on the aforementioned dataset we were able to draw significant conclusions on the dynamics that characterize the sentinel related traffic focusing mainly on well-being aspects that are related with nutrition patterns within the population.

Acknowledgements

Before I proceed with this thesis I would like to express my sincere gratitude and appreciation to my thesis supervisor, Assistant Professor Noha Ibrahim. Furthermore I would also like to thank my thesis co-supervisor, Professor Sihem Amer Yahia. Without her guidance and persistent help this thesis would not have been possible. When the day ends, there is always a single person with which you share your universe, a person that is there to support you whether things go well or not. In this particular case I also owe this person being an budding scientist. Not only because she discovered a talent I didn't know I had back in 2003 but more importantly because she offered me a perspective away from mediocrity. This thesis is dedicated to her, Sofia Kleisarchaki, PhD student. In addition, I feel I need to thank my parents for supporting me at all costs all these years. They were always deeply concerned for my education from a very early age therefore I hope I fulfilled their expectations. Finally I want to thank all my close friends whose support and thought was always present regardless the 1060 nautical miles of sea and land that separate us.

Table of Contents

Illustration Index: Drawings

Illustration Index: Tables

Illustration Index: General Illustrations

1 Introduction

Computer science from its inception until today, has experienced an unprecedented growth unlike any other field. In just a few years, computers transformed from being a privilege of big institutions and governments to a public acquis. Computers size and cost has shrunk over the years while processing capacity doubles every four, providing everyone with the opportunity to own a personal computer. Computers have infiltrated our lives to such an extent as to be an indicator of quality of life. Countries are now judged by the percentage of their people that have access to Internet services while persons that lack computer basic knowledge and skills are considered to be electronically illiterate.

The last two decades we became witnesses of a rapid development of distributed computing and computer networks. Users that were initially restricted to access static text data that was available on the Internet are now enjoying multimedia content that is even produced by other users in real-time. The increasing proliferation and affordability of Internet devices, as well as the ease of publishing, searching and accessing information on the web encourages the individual users to communicate their content with the web society. This gave birth to the radical idea of social interaction over the Internet which in conjunction with the advent of web 2.0 technology led to the appearance of the first large social media. Social media are defined by Kaplan and Haenlein[1] as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of User Generated Content (UGC). The once one way communication with the end users passively consuming web content turned into many-to-many communication of interactive dialogues and dynamic content. These applications are today responsible for a large proportion of the information that are exchanged through the Internet every day. They produce data-sets of massive size that have the tendency to evolve over time.

In recent years, the ubiquity of communication networks speeds up the development of Internet applications. Social networking[2] has been driving a dramatic evolution due to the increasing use of Web 2.0 elements such as blogs, micro-blogging services (e.g. Twitter), social networking sites (e.g. MySpace, Facebook, LinkedIn), social media news (e.g. Digg) and wikis, etc. One of the fundamental problems that emerged in social media stream analysis with a lot of potential applications is to effectively detect underlying topics and their associated documents which are created by the users' interaction and activity. The problem of identifying significant topics or events within a corpus of documents is not new and shares a series of characteristics with two other major and at the same time traditional problems in related literature, Event Correlation and Complex Event Detection. However the encountered scope in our case is different. In a dynamic environment of social networks the network structure evolves rapidly and the content is significantly larger in size. Such observations are unique in the online scenario and challenge the scientific community.

1 A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media” Business Horizons, 2010.

2 B. Buter, N. Dijkshoorn, D. Modolo, Q. Nguyen, S. van Noort, B. van de Poel, A. Ali, and A. Salah1. Exploratory visualization and analysis of a social network for arts: The case of deviantart. Journal of Convergence Volume, 2(1), 2011.

1.1 Motivation & Problem Statement

It becomes clear that modern social services and social media show a substantial potential of providing society with a rather promising source of information. Even though this decentralized information collection and delivery model lacks funding and structural organization, being mainly based on the contribution of non-professional and casual users, this alternative source of information prevails over the traditional news services on a series of important dimensions.

The first significant difference between the traditional information delivery systems and the modern social media is propagation velocity. The plurality of the existing mechanisms rely on clearly defined, complex procedures involving a large number of participating people who possess different expertise. Although this ensures a level sanitization and enrichment and thus content quality, it has a dramatic impact on the time-span between the time an event occurs and the time this event is presented to the world. As it is easily understandable, involved organizations, regardless if they are public or private are restricted in terms funds and human resources and are in a constant quest of controlling their capital or maximizing their profit. This struggle for survival in an unethical and cruel economic world is also reflected to their agenda where different applications compete for recourses leaving less popular applications sacrificed on the altar of capitalism.

This has an immediate consequence on the information coverage as a small finite thematology will make it to the world. On the other hand modern social media and services have virtually unlimited space for news publishing. Moreover private mass media seem to be dependent on large capital funds and publishing firms, eliminating any trace of pluralism. In this unfriendly environment a large fraction of events would be doomed to obscurity without the existence of modern social services. Last but not least, the added value of social media services is the subjective dimension that chaperons news broadcasting, providing an interesting insight on society's opinion on a specific event and the reaction regarding it. These advantages encouraged the academic community to investigate a framework under which, a partial replacement of the traditional sentinel surveillance services with web enhanced ones could take place. This emerging need to provide distributed health informational systems becomes even more apparent if we consider the global economic collapse, which imposes even more severe constrains to the already struggling government agencies.

1.2 Report Organization

After this introductory section, we proceed with chapter 2 where we make an introduction to basic concepts that accompany topic and first story detection on social streams. Having acquired a solid view on topic detection characteristics and dynamics we proceed with an analysis of one state of the art text analysis mechanisms that classify textual content, that is the cornerstone information in every modern large scale social media service. Proceeding with the report, chapter 3 focuses on related work including recent works on the domain of syndromic surveillance using data mining concepts, extracting valuable guidance for our effort. In chapter 4 and 5 we present our workbench and the performed exploitation of the informational wealth of Twitter in an effort to understand the dynamics and the most impacting topics that lie under the veil of the awing social stream. Finally, after the experimental evaluation part, chapter 6, presuming on the knowledge gained from the previous chapter help us extract valuable conclusions on the performance of existing first story detection systems on real-world applications.

2 Theoretical Background

2.1 Social Media

Social media came as a result of the radical, at least for its era, idea of social interaction. As mentioned in the introductory section contrariwise to what was considered as common practice up until this point in history, developments in computer science enabled end users to create, share and exchange information and ideas forming virtual communities. Furthermore, relying on the new highly interactive platforms and the newly introduced web-based technologies social media triggered substantial and pervasive changes to the communication between organization and individuals. As we can understand, the transition from the traditional media to the modern social media became possible due to a series of aspects as shown in the following list:

1. Quality
2. Reach
3. Publishing frequency
4. Usability
5. Immediacy
6. Permanence
7. Involvement

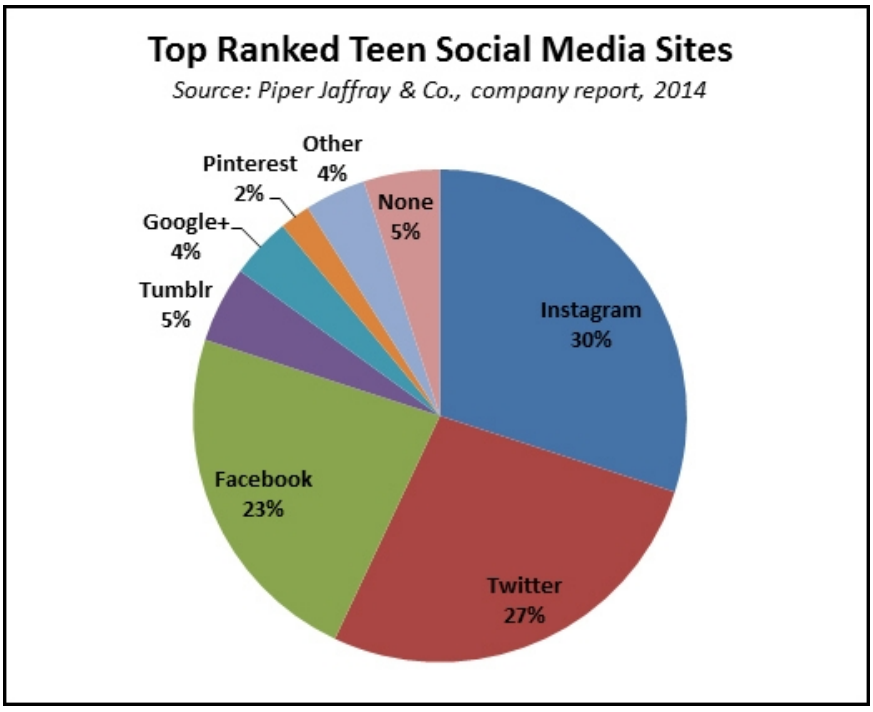


Illustration 1: Top ranked by teens social networks

This revolution in Media benefited the acceptance of society to such an extent that Internet users' interest is nowadays monopolized by social media sites. The extent of this phenomenon is so striking that has introduced significant changes even to related but purely scientific fields like that of Internet Supervision & Measurements by pushing the existing boundaries and provoking the devision of new algorithms that are able to deal with the dynamics of the social media generated traffic. As expected, the aforementioned situation was the ideal environment to incubate a wide range of different types of social media. As fruits of this efforts we can already count more than a dozen different application-wise applications ranging from simple blogs and podcasts to very complex and demanding in terms of maintenance social-networking sites.

Unsurprisingly the academic community was not left untouched by the aforementioned pluralism of social media triggering an effort to provide an official and publicly accepted categorization of the latter ones. Although the revolution of the encountered features often makes the distinction among social media applications to fade, Kaplan and Haenlem published in 2010 a very successful article which was entitled “Business Horizons”. The latter one was based on a series of theories from the field of media Research including among others Social Presence, Media Richness and Social Processes resulting into six different categories of social media as shown in the following summarizing table:

Category	Main Representative
Collaborative Projects	Wikipedia
Blogs & Micro-blogging	Twitter
Content Communities	Youtube
Social-Networking Sites	Facebook
Virtual Game Worlds	World of Warcraft
Virtual Social Worlds	Second Life

Table 1: Modern social media classification

2.2 Data Mining

The key word for everything that concerns data mining is discovery. The latter one successfully summarizes the effort for uncovering interesting, not-obvious information or conclusions from a collection of documents. In reality, data mining goes one step further focusing more on the automated or semi-automated analysis of very large datasets consisting of database tuples or documents in general, all in an effort to derive an interesting distribution or pattern that has never been observed or reported in the past. The discussed field is not new and has a solid presence for a number of centuries already. The main difference with previous applications is now located on the raw material in question itself introducing new dimensions to the problem. As mentioned before, in recent years, social networking has been driving a dramatic evolution due to the increasing acceptance and thus extensive use of the Internet. As an immediate consequence the encountered datasets have dramatically evolved over time and are now characterized by their immense size, complexity, heterogeneity and tendency to evolve over time imposing a number of challenges and restrictions. As we can understand, these emerging conditions made apparent a significant void in related literature. Fortunately enough, computer science

is a discipline whose branches have the tendency to evolve in a parallel way. This permitted the academic community to tackle with the problem of modern data-mining by using advances made in other computer science fields including among others the Neural Networks, Clustering & Genetic Algorithms, Decision Trees and Support Vector Machines.

In the plurality of the cases, the aforementioned task is organized following a multilayer bottom-up approach which starts with the collection of the required data and ends to the conclusions extraction itself, which basically constitutes our main objective. Although variations exist, the plurality of the existing implementations will usually incorporate the following blocks of layers:

1. Data collection
2. Preprocess
3. Data transformation
4. Mining
5. Knowledge Extraction

It is crucial to make clear that this architecture is not in any case binding and those layers may consist of other sub-layers depending on the encountered application. As seen in the previous list, the next block having acquired a sufficient amount of raw data regards the preprocess that needs to be applied to our dataset, process that will render our corpus homogeneous and thus suitable as an input for the following blocks. The importance of the latter one is justified from the inextricable connection between the size of the final corpus and the possibility to extract a pattern. Depending on the application, apart from the aforementioned sanitization process, this layer can also include an enrichment procedure, a step that is gradually gaining ground due to the development of the semantic web.

Bypassing the layer of data transformation which is more procedural and withholds little interest for now compared with the remaining blocks, we usually encounter the blocks that regard the pattern extraction and validation. The latter ones, which constitute the core of the encountered objective in the plurality of applications will incorporate at least one of the following techniques:

1. Anomaly detection
2. Association rule learning
3. Clustering
4. Classification
5. Regression

Starting with the first technique, anomaly detection focuses on the detection of a document or point that significantly diverges from the resultant of the total corpus in terms of a specific measure. Association rule learning on the other hand tries to correlate distributions that are extracted from the same data each one regarding a different variable. Proceeding with the remaining techniques, the next one in our list, clustering can be encountered in a wide range of application besides data-mining. The term clustering describes a procedure that tries to detect underlying physical partitions on a dataset,

with those partitions being intimated by exploiting relevancies that occur among a number of documents or data points in general within our corpus. Classification on the other hand is a process that has bases its operation on the creation of describing data types that consist of a series of required characteristics. After a set of such describers which in this specific case are called classifiers is created, a verdict for whether a document belongs a given category can be provided, a procedure which shares some similarities with the statute of juries in real-life courts. Last but not least is the well-known technique of Regression, a technique bases its operation on knowledge gained from the field of statistical analysis. The main objective of the latter one is to provide a mathematical function that is able to model the documents or data points that consist the encountered corpus with the least error.

Closing, the last step which should always accompany any data-mining task is validation. This is a procedure that is inextricably connected with the credibility of our extracted conclusions. Although a wide range of statistical hypothesis tests do exist in related literature all of them tackle with the same need, the elimination of patterns extracted from previous steps that do not show any substantial scientific value.

2.2.1 Topic & Event Detection

Before providing a definition for the problems of event and topic detection we must first introduce their basic cornerstone. An event can be defined as something that occurs in a certain place and time. Apart from these features, events vary depending on their content in terms of significance and nature of information.

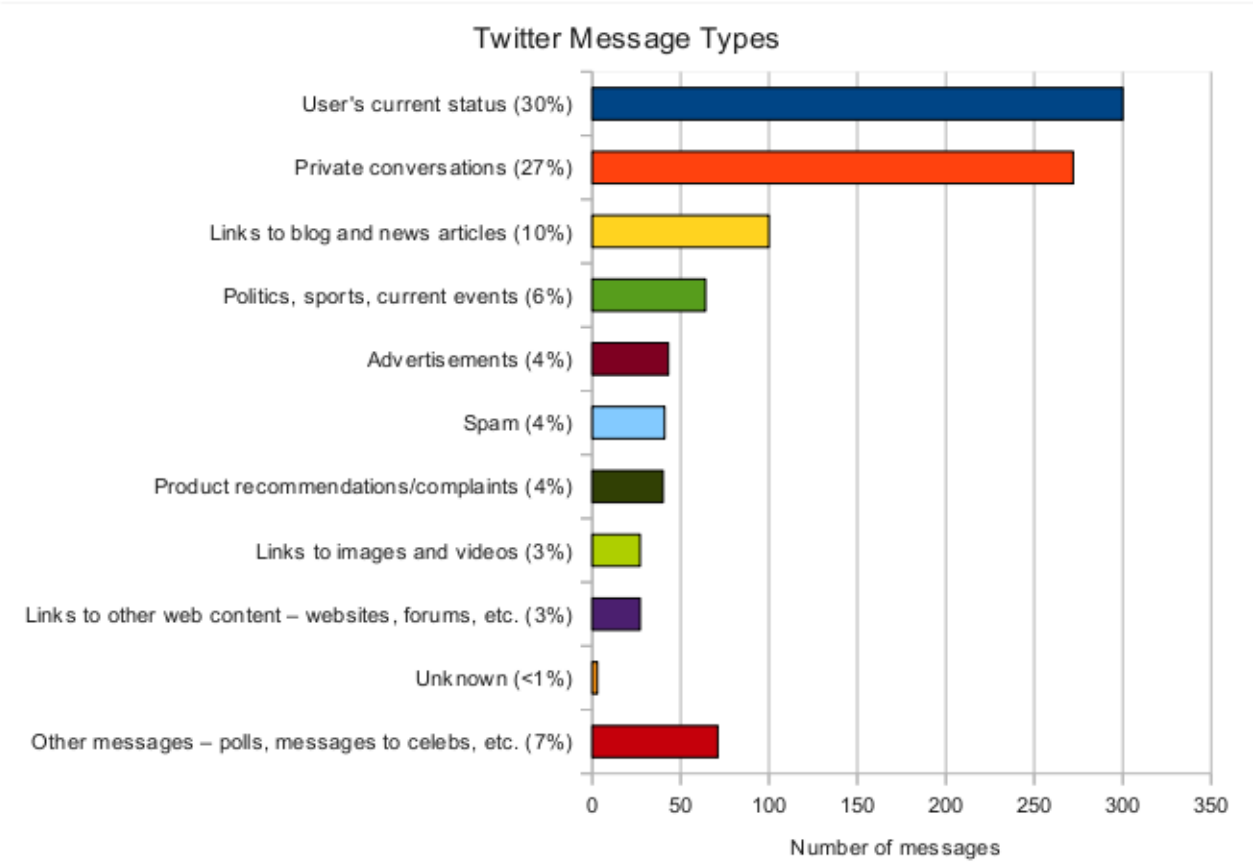


Illustration 2: Twitter messages content

In this context an event detection algorithm should be able, given a corpus that consists of a sufficient number of documents identify underlying events and equally importantly associate documents with their corresponding events. On the contrary First Story detection can be conceived as a specialized subcase of the event detection problem. In this case instead of dealing with static corpuses a first story detection should, given a continuous stream of documents, be able to identify the first story to discuss a particular event. In this context, each arriving document is examined against all older documents. If the new document shows a significant similarity to an already detected story, an association can be safely implied. On the other hand if an arriving document is characterized by a high novelty score we can either consider it as a sign of a new story and create a new cluster or classify it as a minor-importance event.

Event detection does not constitute a new concept that has emerged due to the appearance of the first large social network. Event detection techniques have already been applied on traditional news-wire aiming to discover and cluster events found in textual news articles. The main difference is located on the nature of the examined information. News articles adhere to certain grammatical, syntactical and stylistic standards that make the appropriate for their venue of publication. Moreover due to the nature of printed press most of the published content clearly refers to an existing event negating the need of excessive filtering.

2.2.2 Large Scale Topic Detection

Unfortunately traditional event detection methods fail to adapt to the modern social networking environment due to a series of challenges. This comes as an immediate consequence of the special dynamics of the dataset that are produced by the modern social services. These datasets are characterized by their vast size and their tendency to rapidly evolve over time. Even worse, social network document streams are characterized by a high level of noise and a significant heterogeneity, partly due to the freedom that accompanies modern social services. Taking this into account by the time an event would be identified using traditional news-wire event detection methods all information regarding it will no longer possess any value, forcing us to seek algorithms that will be able to process each document in a finite short time-span.

Authors of [3][3] and [4][4] cast the problems of event and first story detection as a clustering problem on social media documents in an effort to exploit the existing clustering literature. In this context they focused on learning similarity measures in order to produce structures that will allow the use of existing clustering algorithms. As mentioned previously two documents characterized by a high degree of similarity are more likely to be associated with the same event and thus belong to the same cluster. In this context each arriving document must be compared with all other documents.

Intuitively, social media documents clustering is based on social media stream documents manipulation. This manipulation heavily relies on the variety of content features which in their turn depend on the document's data type. As mentioned earlier modern social networks encourage people to publish a wide range of information including photos, videos and simple textual content components. For example an image is characterized by its location while a video can be characterized by its duration. Finally, documents also consist of system generated meta-data including among others timestamps or other geographical information.

3 Hila Becker, Mor Naaman, Luis Gravano: Learning Similarity Metrics for Event Identification in Social Media. WSDM 2010 .

4 Sasa Petrovic, Miles Osborne, Victor Lavrenko: Streaming First Story Detection with Application to Twitter. HLT 2010 .

2.2.2.1 Latent Dirichlet Annotation

Latent Dirichlet Annotation corresponds to a well known tool in the domain of natural language processing. The latter one generally constitutes a model which allows observations to be explained by unobserved groups that explain why some parts of the data are similar. The basic idea behind the algorithm is that each document refers to a small number of topics. Extending this idea, we assume that each word within a document is considered to be attributable to at least one of the aforementioned topics. The aforementioned concept is very similar to probabilistic latent semantic analysis or pLSA. The main difference among these schemas is that in LDA the topic distribution is assumed to have a Dirichlet prior, which in reality permits the model to produce more reasonable mixtures of topics per document.

Furthermore, similarly with documents, extracted topics may consist of more than one semantic content. For example, an LDA model can extract a topic that can be classified as CAT_related while at the same time the same topic can be also considered as DOG_related. A topic has probabilities of generating various words, such as milk, meow, and kitten, which can be classified and interpreted by the viewer as "CAT_related". Naturally, the word cat itself will have high probability given this topic. The DOG_related topic likewise has probabilities of generating each word: puppy, bark, and bone might have high probability. Words without special relevance, such as the, will have roughly even probability between classes.

This behavior is expected as it directly derives frotable:m the way LDA extracts the topics that lie under the surface of an examined dataset. In this direction, it is crucial to assimilate that a topic is not strongly defined, neither semantically nor epistemologically. On contrere, the whole process is mainly based on co-occurrence of words. More specifically, a topic is identified on the basis of supervised labeling and manual pruning based on the likelihood of co-occurrence. The latter is easily conceivable by realizing that a term may participate into numerous topics with a different probability, however, with a different typical set of neighboring words in each topic.

2.2.2.2 Enhancing Topic Detection

Despite its advantages and being proved efficient in dealing with corpuses that consist of large text documents, Topic Detection Algorithms has one significant drawback. It does not take into consideration any syntactical characteristics nor any conceptual or semantic content that may be implied by the examined documents. Moreover the exclusive employment of the LDA model also proved to be inadequate for processing social stream documents, mainly due to their previously mentioned restricted size and linguistic diversity. In an effort to improve proposed system's performance a series of text processing steps usually involved in inverted indexes building can be applied selectively including:

- Stemming
- Stop words elimination
- Dimensions Minimization

3 Related Work

3.1 Examining the benefit of exploiting social media data

As expected, the public health community before applying data-mining techniques on social media started with a task that proceeds semantically the latter one and is no other from the examination of the value that chaperons the social media derived knowledge. In this context the academic community had to consider how social media can be used to spread health information, with applications including risk communication and emergency response. In this direction, Vance, Howe, and Dellavalle (2009)[5] analyzed the pros and cons of using social media to spread public health information in young adults. Pros include low cost and rapid transmission, while cons included blind authorship, lack of source citation, and presentation of opinion as fact. Greene et al. (2010)[6] studied how medical information is exchanged on Facebook, where groups specific to diseases share information, support, and engage patients in their diseases. Fernandez-Luque, Karlsen, and Bonander (2011)[7] reviewed different approaches for extracting information from social web applications to personalize health care information. The model we use in this paper could be used to analyze tweets for health care personalization. Finally, the community is considering the larger impact of how social media can impact health care, where patients can "friend" doctors and constantly share information among thousands of friends (Hawn 2009; Jain 2009)[8][9].

3.2 Epidemic-Centric

For reasons that have been thoroughly presented in previous chapters of this report, the Web has become a source of syndromic surveillance, permitting us to operate on a wider scale requiring significantly reduced resources. As expected, due to the novelty of the idea, the academic community turned to candidate ailments that are characterized by the highest success possibility. Soon, it was clear that ideally the selected ailments should have epidemic-centric characteristics make them easily identifiable from other illnesses. One of the first works in this direction is Google Flu Trends which was published by Ginsberg et al. Back in 2008[10]. The latter one is very interesting as it does not rely on social media derived data but tracks the rate of influenza using google query logs on a daily basis. Experimental evaluation later proved that this system is up to 7 to 10 days faster than the Center for Disease Control and Prevention's (CDC) FluView, a system that was conceived by Carneiro and

5 Vance K, Howe W, Dellavalle RP. Social Internet sites as a source of public health information. Dermatologic Clinics, 2009 Apr;27(2):133-6. PMID: 19254656

6 Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen, Intern Med. 2011 Mar;26(3):287-92. doi: 10.1007/s11606-010-1526-3. Epub 2010 Oct 13.

7 Fernandez-Luque, L.; Karlsen, R.; and Bonander, J. 2011. Review of extracting information from the social web for health personalization. Journal of Medical Internet Research 13(1)

8 Jain, S. H. 2009. Practicing medicine in the age of facebook. New England Journal of Medicine 361(7):649–651.

9 Hawn, C. 2009. Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. Health Affairs 28(2):361–368.

10 Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014.

Mylonakis 2009[11]. As expected, the success of the aforementioned work intrigued others to examine the potential of query logs usage in favor of syndromic surveillance. In this direction shortly after the pioneer effort we encounter a set of works published by Pelat et al. and Seifter et al. respectively[12]. Although these works differ in terms of application, as the first one tries to correlate Google queries with a series of diseases and the second one is more specialized focusing on "Lyme" disease, both works fall under the area of infodemiology as defined by Eysenbach in 2010[13].

Moreover, similar results exist for Twitter, which can be a complimentary resource to query logs, but also benefits from an exceptional informational wealth that is freely available from the users that communicate their content with society. In this context, another series of works are found in related literature. Lampos and Cristianini in 2010[14] and Culotta in 2010[1516] too, managed to successfully correlate tweets mentioning the flu and related symptoms with historical data. On the other hand, Quincey and Kostkova (2010)[17] collected tweets during the H1N1 pandemic for further analysis, work that followed a work from Ritterman, Osborne, and Klein back in 2009[18] that combined prediction markets and Twitter to predict H1N1 outbreaks. Lastly but equally interestingly, following a more general approach, Chew and Eysenbach (2010) evaluated Twitter as a mean to monitor public perception of the 2009 H1N1 pandemic.

3.3 Other Ailments & Behavioral Risk Factors

Up until now we have briefly examined a series of related works that focus on ailments that show a clear epidemic behavior with data mining from Twitter and other social networks being mainly concentrated in tracking trends of the virus of influenza. As mentioned numerous times in this report the latter one is considered to be ideal for such processing since its geometric flood-like spreading pattern is easily identifiable among other ailment episodes which lay on a time-line. A first attempt to change this literature status quo was carried out by Michael Paul and Mark Dredze in their work "You Are What You Tweet: Analyzing Twitter for Public Health"[19]. The latter one which is presented including more details due to the connection with our thesis, focuses on three main objectives, Syndromic Surveillance, Behavioral Risk Factor Analysis and Medication Usage Analysis which share the same starting point as all three rely on the same data-mining tool, Topic Detection. The need for an algorithm that will be able to extract a series of different topics is justified by the nature of the dataset itself, which consists of a very large number of documents that are expected to belong to numerous ailments or other health-being topics.

In order to generate valuable knowledge from the enormous number of unstructured micro-

11 Carneiro, H., and Mylonakis, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–64.

12 Pelat, C.; Turbelin, C.; Bar-Hen, A.; Flahault, A.; and Valleron, A.-J. 2009. More diseases tracked by using google trends. Emerg Infect Dis 15(8):1327–1328.

13 Chew, C., and Eysenbach, G. 2010. Pandemics in the age of twitter: Content analysis of tweets using the 2009 h1n1 outbreak. PloS ONE 5(11):e14118.

14 Lampos, V., and Cristianini, N. 2010. Tracking the flu pandemic by monitoring the social web. In IAPR 2nd Workshop on Cognitive Information Processing (CIP 2010).

15 Culotta, A. 2010a. Detecting influenza epidemics by analyzing twitter messages. arXiv:1007.4748v1 [cs.IR].

16 Culotta, A. 2010b. Towards detecting influenza epidemics by analyzing twitter messages. In KDD Workshop on Social Media Analytics.

17 Quincey, E., and Kostkova, P. 2010. Early warning and outbreak detection using social networking websites: The potential of twitter. In Electronic Healthcare. Springer Berlin Heidelberg.

18 Ritterman, J.; Osborne, M.; and Klein, E. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In Workshop on Mining Social Media.

19 Michael J Paul, Mark Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. ICWSM 2011.

blogging documents the authors of the examined paper used the recently introduced Ailment Topic Aspect Model. As we have already mentioned before, this model shares many similarities with the Latent Dirichlet Annotation model which is also described in chapter 4. ATAM and its extension ATAM+ models proved to be ideal tools for the specific task as they both proved to be adequate for the association of each document with an underlying ailment or other well-being topic. ATAM+ was the result of an effort to enhance system's performance. In this context provisions have been taken to incorporate previous knowledge in the form of a credible external source and more specifically healthcare articles. These healthcare articles which are treated as ground truth during the methods training phase proved to have a positive effect on the extracted knowledge in terms of a set of measures including among others Pearson Correlation, MRR and other statistics. From this point the authors were able to apply a different technique on the tweets of each topic as they see fit accordingly to the three main objectives.

Regarding Syndromic Surveillance, the examined work initially followed the beaten path, trying to track influenza events over the time variable. In order to evaluate the extent to which results correlate to the actual influenza trends, results are compared to the official data that are published in a weekly base from the Center for Disease Control through the FluView program. After a positive quantitative evaluation of the experiment results, Michael Paul and Mark Dredze went one step further, as they tried to track another ailment on a different axis, location. For this purpose the tweets that refer to allergies were isolated and furtherly pruned in order to correspond to the registries that took place within the United States. Interestingly, results showed clear patterns that match what is publicly known for a number of states. Lastly, exploiting the fact that the CDC not only published the time of an event but also the location a third task of Syndromic Surveillance was performed, using this time both axis succeeding a correlation of more than 85%.

Returning to the second challenge authors faced, a significant effort has been put into identifying potential hazardous behavioral factors for different geographical regions within the US. For each statistic that could potentially be measured with one or more ATAM/ATAM+ ailments, authors measured the Pearson correlation coefficient between the ailments discovered in each US state with the state's risk factor rate. Considering the difficulty of the whole venturing positive correlations have been found among what was documented by the CDC via telephone campaigns and what was discovered from the two employed models.

Regarding the last objective, Medication Usage Analysis, work tried to shed some light to immune ailments that would probably pass unnoticed from the public healthcare. Furthermore the system also provides some indication on the perception of society on drugs usage providing us with some very interesting observations for specific categories of medications. The second aspect has been also examined in the past by Scanfeld, Scanfeld, and Larson in 2010. The latter ones evaluated the public understanding of antibiotics by manually reviewing Tweets that showed incorrect antibiotic use, e.g., using antibiotics for the flu.

Summing up, despite the publicly recognized drawbacks that characterize datasets that derive from twitter's social stream, including the blind authorship, the lack of source citation, the heterogeneity, the noise and the presentation of an opinion as fact, authors believe that Twitter can have a great impact on public health informatics providing a reliable alternative for sentinel surveillance agencies that will or already run on a tight budget. In order to support their case, authors have investigated a variety of public health data that can be automatically extracted from Twitter. Although results do justice to their case, however twitter cannot and must not be treated like the jack of all trades as it may not be reliable for certain types of information. This motion of censure can be justified if we taking into account, the special nature of social networks' demographics in general, the low availability and accuracy of geographical information and the low number of tweets per user. Unfortunately, in the

health related messages, 71% of the users have only one single tweet and 97% have less than 6, a percentage insufficient for computing user level statistics. Fortunately, further adoption of mobile devices by a larger part of the population will probably provide us with more detailed geographical information. In conjunction with the steadily increasing popularity of Twitter outside the US the boundaries of this research direction might re-estimated sooner than expected.

3.4 Nutrition Oriented

Apart from the works that regard health aspects of our life, a series of works do exist that regard nutritional aspects, positioning them close semantically to our interests. Among them we distinguished a work published by Kyle W. Prier , Matthew S. Smith , Christophe Giraud-Carrier , and Carl L. Hanson in 2011 and was entitled “Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic”[20]. The selection was not only based on the content itself but also because this publications clearly shows the capacities and on the other hand deficiencies of applying the LDA algorithm on huge conversational datasets from Twitter, partly justifying the limitation we encountered within this thesis.

Through this empirical study, the authors directly addressed the problem of how to effectively identify and browse health-related topics on Twitter. This task, was carried out by applying Latent Dirichlet Annotation on a two-millions dataset that was collected over a two-weeks period mainly focusing on tobacco usage within the united states. By focusing on this test topic throughout the study the effectiveness of LDA to learn more about health topics and behavior on Twitter was explored. Initial conclusions drawn by the application basically do tend to match the algorithm's definition as the latter one proved to adequate for identification of high frequency topics on large datasets such as weight loss, or topics relating to other governmental healthcare agenda. Interestingly, LDA was able to identify a topic that was characterized by high frequency of marijuana-related terms indicating a potentially significant behavior risk that can be detected through Twitter. While this applying LDA on the whole corpus did not detect lower frequency topics, it is believed that it may still provide public health researchers insight into popular health-related trends on Twitter if a way is found either to “tweak” LDA or provide the algorithm with a more appropriate dataset. As a more practical solution to overcome this problem, authors examined the use of a non-automated method which relies on the careful selection of key terms.

In the case of tobacco use, the results indicated this method may be a valuable tool for public health researchers. Based on the results of LDA topic detection model, Twitter has been identified as a potentially useful tool to better understand health-related topics, such as tobacco. Results suggest that chronic health behaviors, like tobacco use, can be identified and measured over shorter periods of time. However, as the examined work is restricted to tobacco use, no guarantees can be provided for the adaptability of this method to different thematology.

20 Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. In Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction (SBP'11), John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai (Eds.). Springer-Verlag, Berlin, Heidelberg, 18-25.

4 Workbench

As mentioned earlier, in recent years, social networking has been driving a dramatic evolution due to the increasing use of Web 2.0. The major impact of Online Social Networks (OSNs), such as MySpace, Facebook and Twitter on Information Technology, paved the way for popular Internet applications and attracted hundreds of millions of users heavily influencing our everyday life. As expected, this evolution did not leave unmoved the academic community. Contrariwise, this turning point in the history of informatics not only attracted the true interest of the academic world but also and prompted several researchers to examine traffic synthesis, user activity, evolution, and structure, of popular online social networks and services.

Recent efforts on this direction exhibit a very interesting variance in terms of application or general direction. If we want to provide a general outline for two general categories that summon up the plurality of current works and trends, we could briefly classify current works in either those that examine and exploit the underlying network that arises from explicit or implicit connections among different type of nodes, or in those that try to draw useful conclusions based on collective analysis of user generated data which noways includes plain text, photos, moving pictures and other multimedia files. Our work, which mainly belongs to the second group was influenced by a series of recent works that try to extract health oriented knowledge. Although such applications experience their infancy era at the moment, and although network administrators tend to hide such events from public view, as much as possible,all indications confirm the added value that accompany health data that is extracted from social networks and services as it was made clear in chapter two, which regards the state of the art related works.

In an effort to contribute in the aforementioned exertion of the academic community to exploit all the information wealth that derives from social networks for the benefit of the society, we collected a quite significant number of micro-blogging documents originating from the extremely popular Micro-blogging service named Twitter. The latter is considered to be an excellent choice which was based on a series of important factors which start to become clear through recent works in the domain of data mining. Those key factors are quite diverse and range from nature oriented ones, which concentrate on the structure of the exchanged information to more technical ones that focus on the on the availability of data or the maturity of the existing APIs.

Cumulative, over 200 millions tweets were randomly collected over a period of more than two month and they were analyzed accordingly. Although this amount of data sounds imposing, in reality it represents only a small fraction of the user generated content for the same period, and more specifically about the one centime of the total traffic. Although the plurality of the content never makes it to plain sight and in an effort to see the glass half-full it is important to understand that regardless of its restrictions, Twitter is one of the very few social networks that enables researchers to collect user generated traffic along with all their accompanied attributes and to analyze them, deriving interesting results for the network.

Apart from characterizing generated traffic based on the collective analysis of individual tweets, we go one step further, trying to derive interesting conclusions by extracting topics that lie under the surface of the related textual content using the well known in the domain of text analysis, Latent Dirichlet allocation model.

4.1 Tweets Collection and Methodology

As mentioned before, Twitter is a popular social network counting several hundreds of millions members who collectively are responsible for a considerable amount of the information that is shared through the world wide web. For each user and each micro-blogging post posted by the latter one, Twitter allocates a numeric ID in a sequential fashion. This mechanism, which is documented in [21] proved to be critical for our work as apart from defining which tweets will be retrieved through the stream examination, it also permits us to examine traffic that was generated in the past.

Another important characteristic of Twitter that facilitates its analysis is connected with its high popularity within the world wide web society. Foreseeing that the network's high adoption will inductively intrigue independent developers and software houses to implement all sorts of third party applications, Twitter's administrators provide a very solid base in the form of the very well documented Twitter REST API[21].

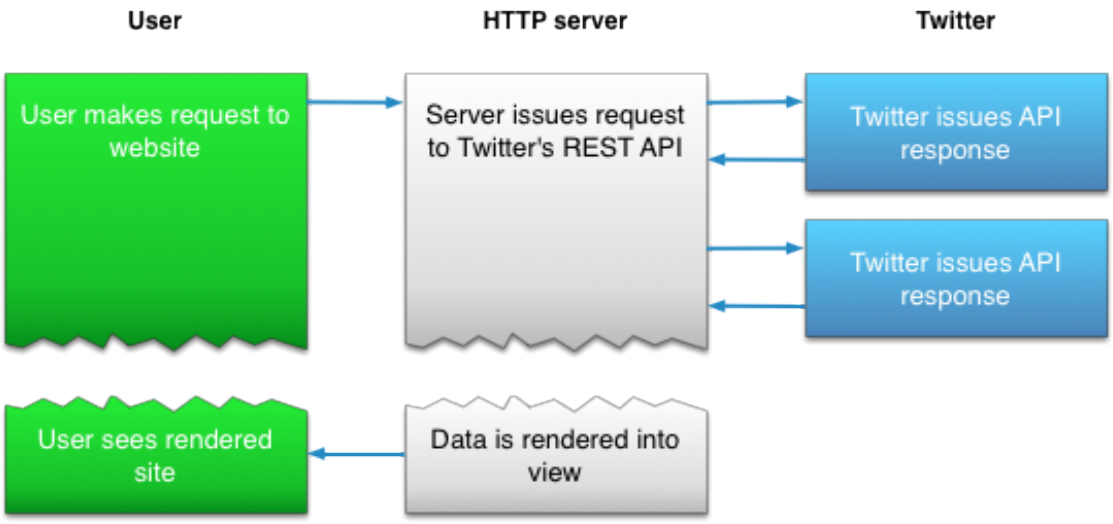


Illustration 3: Data collection layer

Besides the officially provided API, numerous specialized APIs exist, each one focusing on a different platform or programming language permitting end users and developers to write their own scripts and programs enabling them to dynamically collect random user profiles and tweets according to its needs. The aforementioned insfrastructure uncouples us from the need to rely on the availability of enormous static datasets, released periodically by networks' administrators, a policy that is adopted by other equally popular networks like Facebook. In our case, due to its maturity and orientation Twitter4j[22] was selected. The latter one is a cross platform API for the very popular object-oriented JAVA language.

Regarding the more technical aspect of the retrieval, Twitter4j in accordance with the equivalent modules of the official API permits authorized developers to access published data with two different mechanism, the “Stream” and the “Search” APIs respectively. The “Stream” mechanism could be easily

21 https://dev.twitter.com/docs/api
22 http://twitter4j.org/en/index.html

described as tapping the network's output. The only difference is that due to necessary restrictions imposed by scalability problems we can only “hear” about one percent of the user generated traffic. The reality is that even if Twitter's administrators provided us with the total traffic it would be impossible for us to store or process this information in real time due its immense size. “Search” API on the other hand, as its name denotes permits users to construct and execute queries regarding users, posts and all the information that accompany them.

Focusing on the sought information, each user profile records a number of information regarding the user. Apart from the numeric user ID, each user profile is accompanied with a very large and diverse set of attributes including among others:

- Username
- Registration Date
- User's Preffered Language
- User's Location
- Number of Followers
- Status

In a similar approach, apart from the user generated information that users attach to each one of their post, Twitter encapsulates more information to each tweet. Those information include among others:

- Creation Date
- Number of Retweets
- Username
- List of Contributors
- Geographical Information

Although both lists are indicative of the wealth that is encapsulated in the generated traffic it is far from being complete as this is not in our scope of interest. More concrete technical information on the way that Twitter4j models information can be found in [21]. As expected our application requires only a smaller subset of the total available information that concerns our health oriented interests. In this context we intentionally restricted the parsed information in the following attributes:

- Tweet ID
- User ID
- Tweet Creation Datetime
- Tweet Location
- Tweet Text Body

4.2 Alternative Sources

Besides the large number of tweets that were collected using our implementation and the offered by Twitter4j mechanisms we were offered by the authors of [19] a very interesting health oriented dataset. This dataset consists of 1.620.000 tweets and is a stripped version of a huge non-filtered corpus that counts more than two billions tweets. In order to pass all this data through a sanitization step which would eliminate noise, non-english and not health related tweets authors submitted the aforementioned data to a binary linear SVM machine.

5 Statistical Analysis

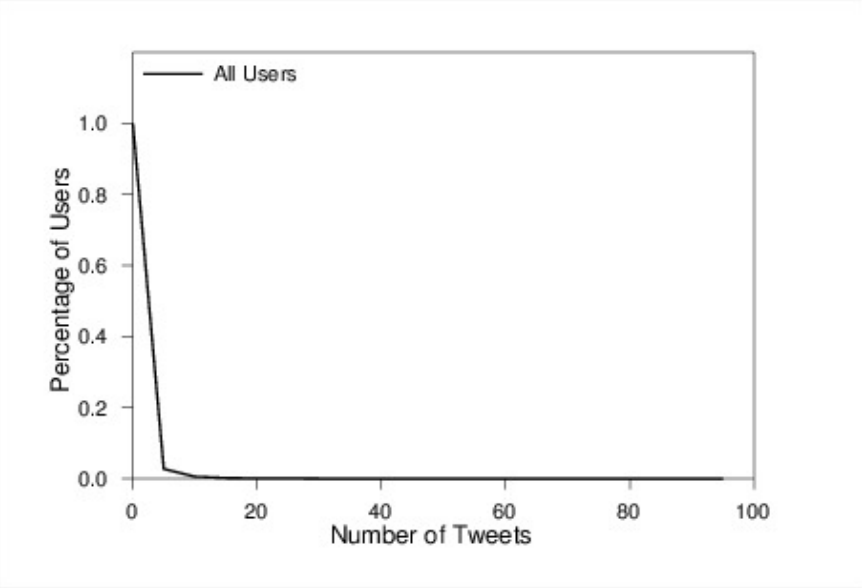
5.1 Dynamics of the human-generated traffic

For both datasets we tried to examine as thoroughly as possible all the extracted tweets investigating their connection with contexes related to health, lifestyle and nutrition. In order tackle with this task we organised our examination on four main axis each one concentrating on a wide range of different variables as follows:

- Users Activity
- Tweets Structure
- Specific Ailments Dynamics
- Other Corpus Statistics

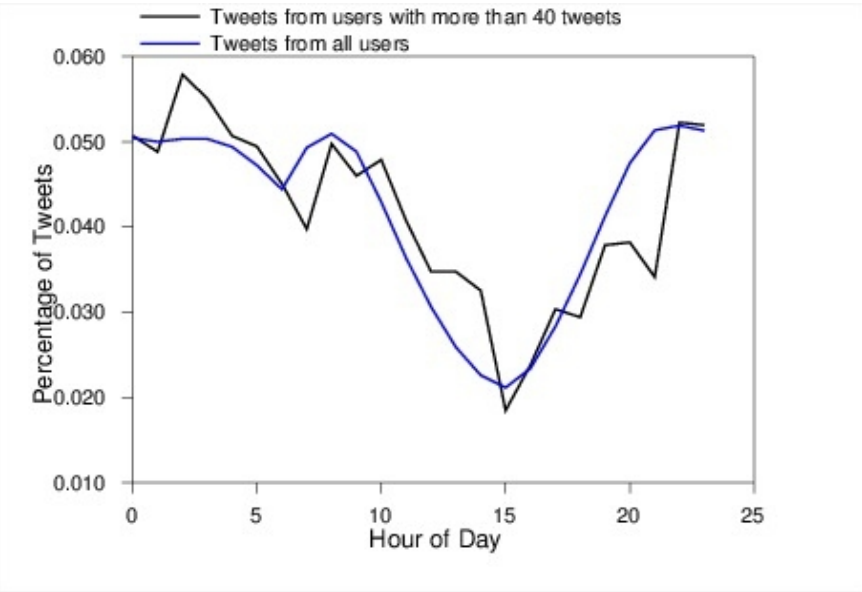
5.1.1 Users activity

In order to examine our corpuses dynamics in terms of activity we focused on two basic questions, how often and when do users express their content as shown in the following figures. Figure [1] although it is one of the simplest from the ones we include in our report is very important as it helps as validate that an already known property that characterizes social networks implies. As shown in the cdf plot in question, the number of tweets per user follows a long tail distribution. This means that a very small fraction of the users that contribute in the corpus is responsible for the largest part of the tweets that participate in the latter one. It is very characteristic that less than 10% of the users have more than 5 tweets within the corpus.



Drawing 1: Number of Tweets CDF

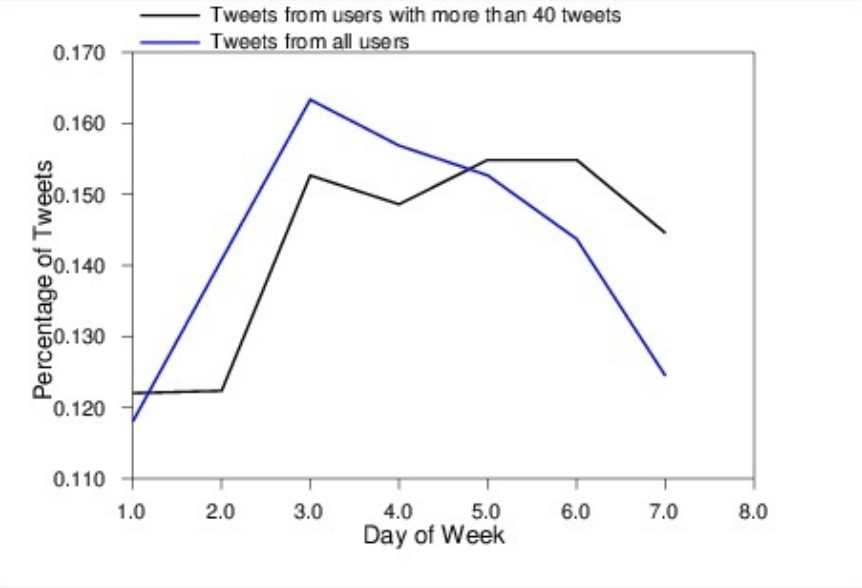
As mentioned before, the second variable concerning users' activity we tried to examine is time. In an effort to locate the time conditions that trigger users to communicate their content and recognizing that specific ailments and well-being related issues are tightly connected with time we proceeded by plotting three distributions, each one focusing on a different timescale. For all following figures regarding time we plotted two separate distributions each one tracking a different users target groups in an effort to examine if the there is any connection between the time instances that tweets are sent and the total traffic that is generated by each user. In this context we organized two groups, one that includes all the users that participate in Dredze's dataset and another one that consists of all the users that have at least 40 tweets within the corpus as we consider those users to be active.



Drawing 2: Distribution of health related tweets round the clock

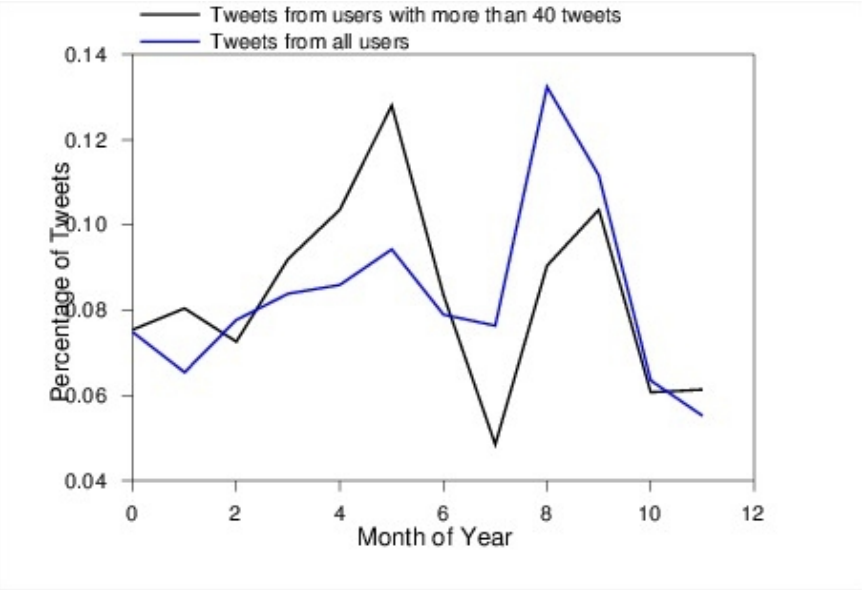
As seen in figure [2] both distributions share very similar tendencies dispelling the hypothesis

that a correlation between time and number of tweets exists. As we can see, tweets posting shows a steady activity with a very clear overall low in 15:00. In general a clear absence with a decline that starts in 10 pm can be observed, fact that could be related with the average working hours in the western world.



Drawing 3: Distribution of health related tweets during the week

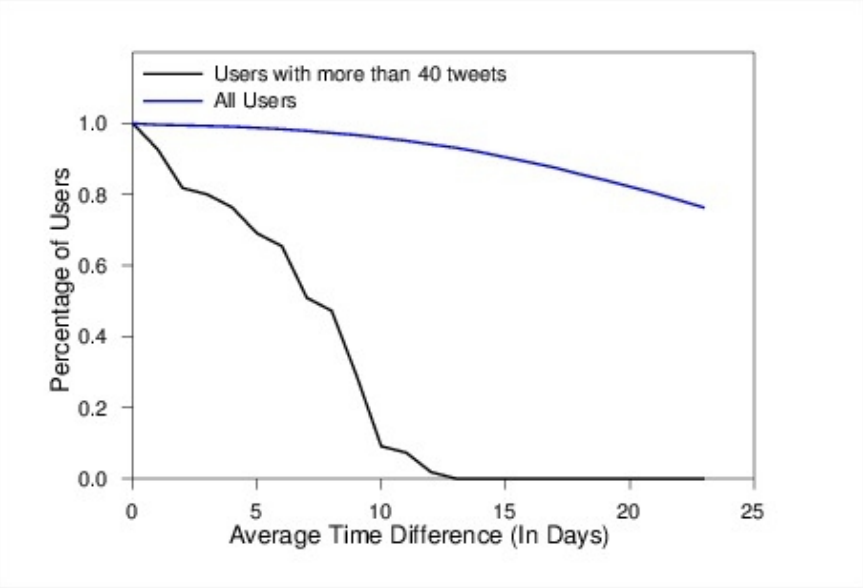
In figure [3] distributions tend to match with a small shift up to a critical timestamp, with different tendencies being followed after this key point in time. In general both categories of users tend to concentrate their activity within the working days of the week. This behavior is so clear that makes us believe that the psychological burden that is added by the intensity of the pace of modern life plays a crucial role in the moment and rate when users choose to post tweets. Another observation that supports this hypothesis is that the peak in both distributions is located in the first days of the week and more specifically in Tuesday. After this day for the plurality of users starts a rather linear decline until weekend is reached. As mentioned before after Wednesday distributions follow different patterns. In this direction active users' distribution differentiates from the second one as the latter ones maintain a high activity and at the same time a second local high exists around Friday, an interesting pair of observations that requires further examination.



Drawing 4: Distribution of health related tweets round the year; Zero corresponds to January

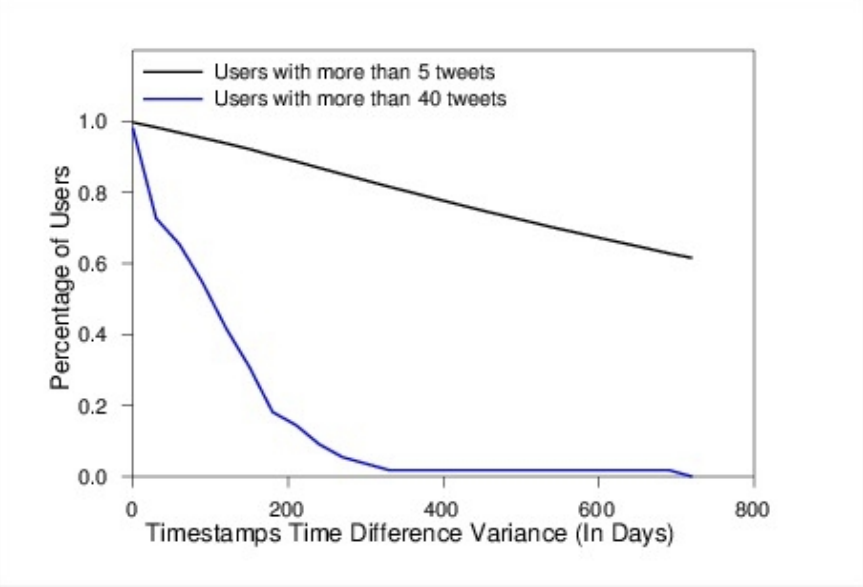
Figure [4] outlines the distribution of the tweets round the year in an effort to examine if there are specific months that concentrate large fractions of the generated traffic. The main hypothesis behind this plot is that within the year we can define two discrete impacting periods for public health. One is located in spring period and is connected to allergic outbreaks while a second one exists in months that are characterized by significantly lower temperatures. As seen in the preceding plot, both distributions equally show similarities but also important differences. Although distributions show similar tendencies in different scales, overall peak periods are different. For active users words period is in May while users with more sporadic presence tend to send the plurality of their tweets within the winter.

Apart from examining the critical points in time that characterize the generated traffic we tried to examine the time variable from another aspect. Fueled by the outbreaking nature of certain health oriented issues we tried to examine if tweets follow similar tendencies. In this direction we extracted the time differences that occur between all tweets that participate in the health oriented dataset. As expected from the following figures we had to exclude users whose presence is restricted to a single or a very small number of tweets, altering the one of the two aforementioned categories. Instead of forming a group that contains the totality of the users that can be found in the discussed corpus we now introduce a new category that focuses on users that have sent more than 5 tweets as it is impossible to derive time differences for a single tweet. The first of the following figures, figure [5] consists of a CDF distribution of the average time difference between the tweets of each user, with axis x corresponding the average time difference for each user, counted in days and y axis corresponding the users' percentage. As we can see plot makes it crystal clear that active users are characterized by a higher sending rate in contrast with the users participating in the more general group whose activity follows a linear uniform distribution in terms of average time difference.



Drawing 5: CDF for the timespan between the first and last tweet per user

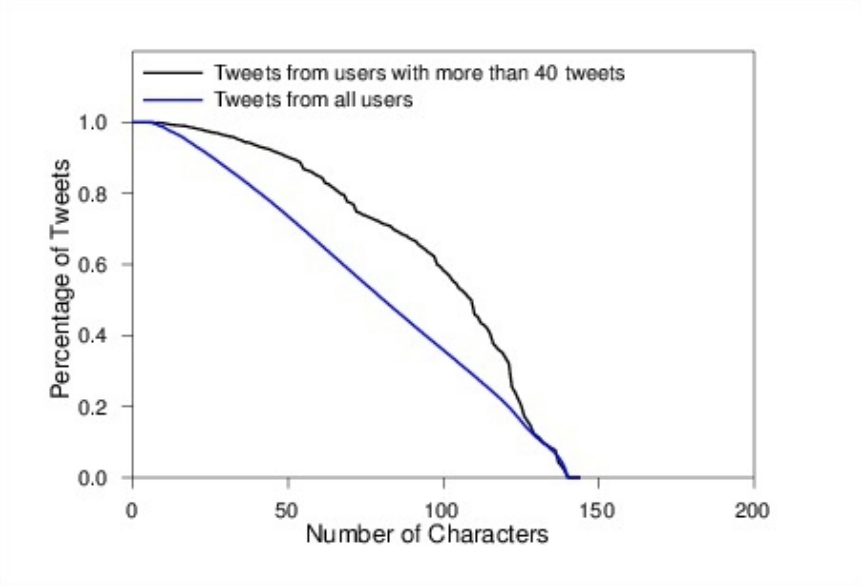
Figure [6] on the other hand examines the same aspect of time differences from a different point of view, which is the standard deviation of the aforementioned time-gaps. While y axis remains practically the same, x axis now represents the average time-gap variance calculated in days. Again, active users tend to have a more steady sending rate while variances values for the plurality of the users population follow a linear uniform decline as variance values become greater.



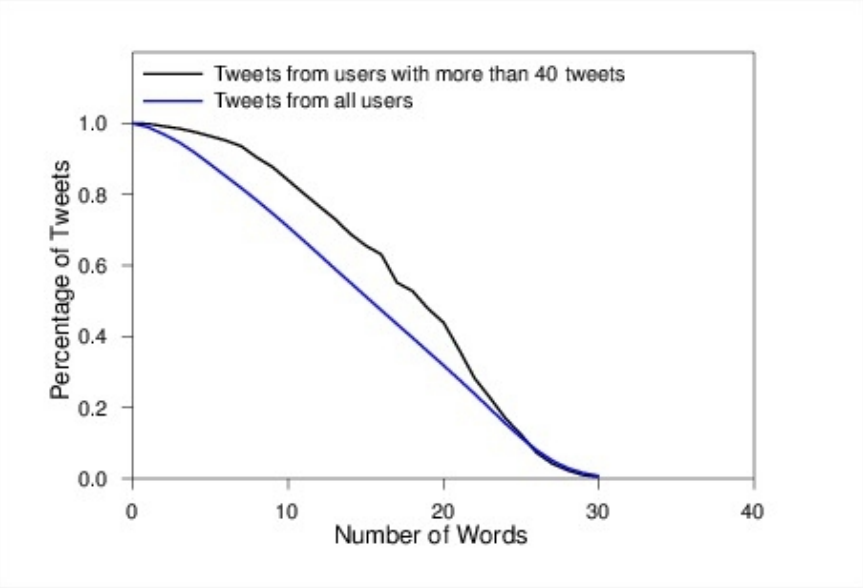
Drawing 6: CDF of the standard deviation of the timespan between the first and the last tweet per user

5.1.2 Tweets Structure

Having acquired a set of initial indications on users' activity and taking into consideration our ultimate target which is to investigate the issues that concern Twitter's community we proceeded with examining the structural nature of our traffic's basic structural unit, tweets. This is a vital prerequisite as users' well-being variables are being imprinted with the help of Latent Dirichlet Annotation technique, and the latter one bases its operation on exploiting the textual content and characteristics of the corpus. Our first concern which is imprinted in figure [7] regards the length in characters for each posted tweet. Equally with figures that were presented earlier, we once more organized our samples on two categories of users. Those categories as before contain the totality of users and the subset of the users that have a presence that counts more than 40 tweets respectively.

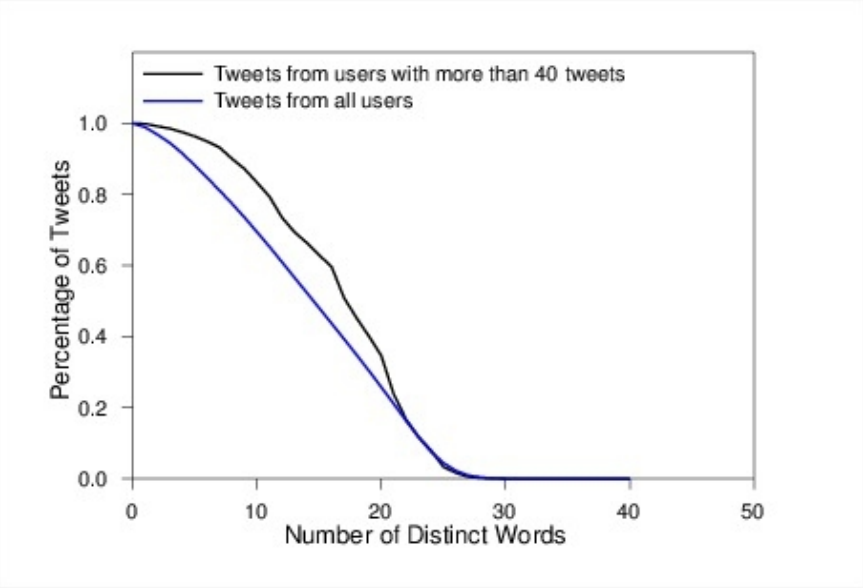


Drawing 7: CDF of the number of characters per tweet

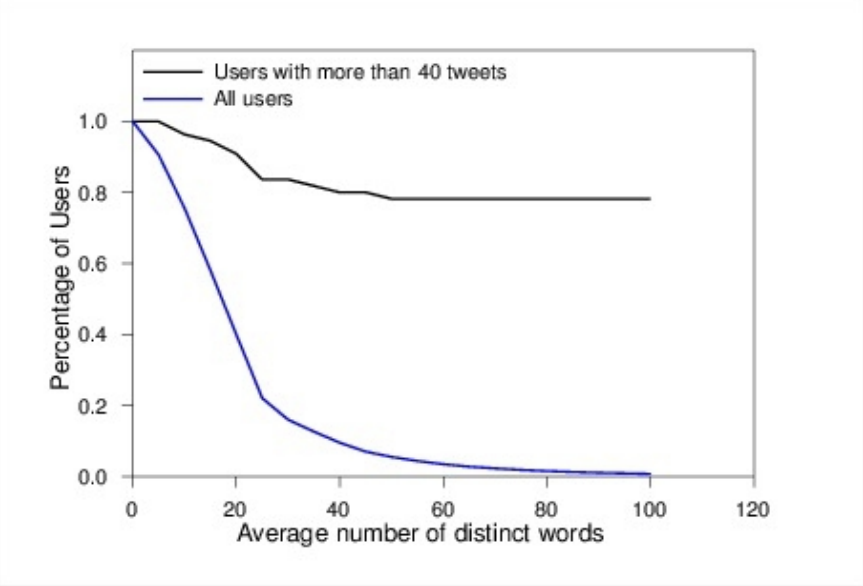


Drawing 8: CDF of the number of words per tweet

As we can see, no tweets with more than 140 exist confirming one of the network's fundamental properties. Moreover we can make to basic observations. The first one is that the namely 'active' users tend to send longer tweets while the second one is related to the distribution of the tweets that come from the totality of the population, which is linear. In an effort to expand the boundaries of what we know about the structure of the tweets we also considered important to see if active users also tend to use a higher number of words per tweet. More specifically 51% of our population has sent an average of more than 30 words per tweet. In the case of the active users this percentage is significantly higher as it passes 65%. Similarly with the previous results, in figure [8] we ascertain that active users not only tend to send longer tweets but also tend to include a higher number of tokens in them. This characteristic did not surprise us as we could consider it as an immediate consequence of the observations made in figure [7]. An indication that can be inductively extracted is that active users probably do not use complex words and do not differ in terms of vocabulary from other users in terms of complexity.



Drawing 9: CDF of the number of different words per tweet

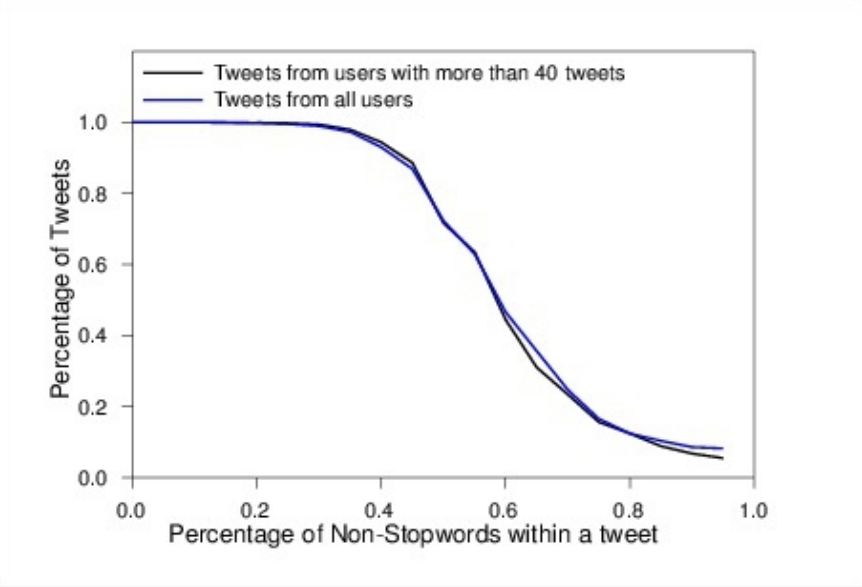


Drawing 10: CDF of the average number of distinct words per user

Observing that often tweets tend to show a repetition of the participating tokens, we proceeded with examining the number of unique words as distributions over tweets and users respectively. Results, which are plotted in the following figures suggest that the previous indication that is based on figure [8] remains valid even after restricting our samples to only distinct words. More specifically, as figure [9] denotes, although both distributions follow a decline and none of the user groups has more than 30 distinct words, the distribution that is extracted by the active users clearly outlines a larger surface. As clearly shown in the following figure, if we take into consideration the total distinct vocabulary per user the aforementioned indications become even more significant and easily identifiable. It is indicative that only 20% of the total population has a vocabulary that counts more than 25 words in total. At the same time almost 80% of the active users has used at least 100 words to

express well-being aspects of their life. Again, although this tendency was expected as it goes hand in hand with the higher number of distinct words, its extent goes beyond our expectations.

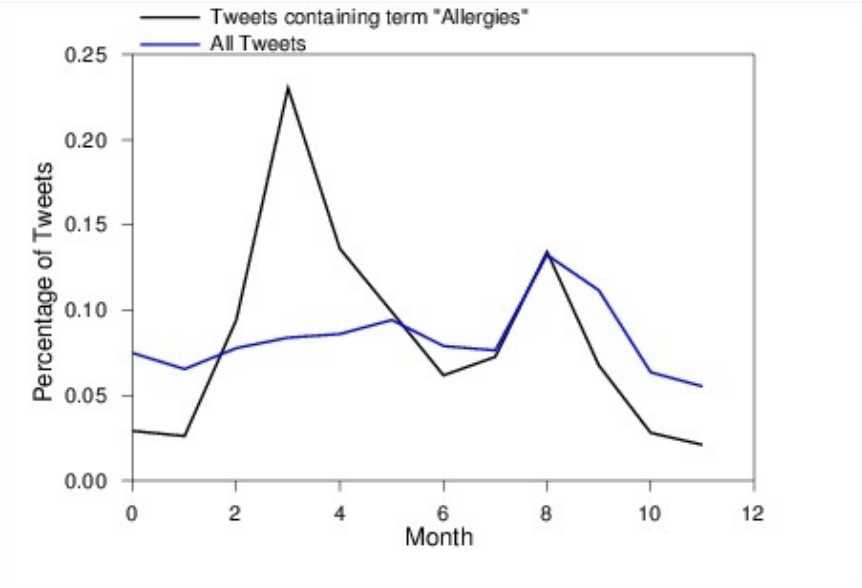
Lastly, we considered as equally important to the previous efforts to examine the discriptive power of the encountered user generated documents. In this context a non-stopword token is thought to provide more clues over a user's state than a stopword. In the following figure the distribution of non-stopwords percentage is plotted for tweets over the time variable. From what we can observe based on the two distributions that are imprinted in figure [11], practically no difference exists among the two user groups. Quantifying the plot in question, only 70% of the generated traffic has less than 50% of stopwords.



Drawing 11: CDF of non-stopwords per tweet

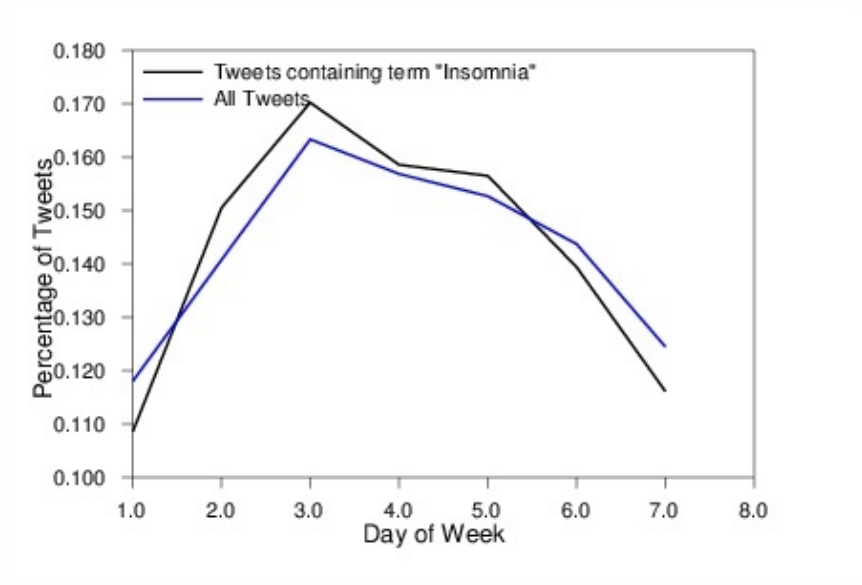
5.1.3 Specific Ailments

Having accuired an initial estimation on the generated traffic's dynamics, we proceeded with examining distributions that correspond to a series of specific ailments. This not only permitted us to explore their impact on our population but also provided a qualitative validation for our corpus extraction methodology. In this direction we tried to select ailments that have a clear and publicly known and accepted pattern. Among the wide range of candidate ailments we selected those that relate to allergies and sleep disorder problems. The first one should have a very clear distribution over the year while the second one should peak within specific time-windows if corpus has been properly collected. In the following figure we plotted the distribution of around 45.000 tweets that correspond to allergy problems in an annual basis. As we can observe, the following figure confirms the hypothesis that allergies mainly occur within March and May while at the same time it reveals a second period when our population is defrayed by allergy outbreaks.



Drawing 12: Distribution of tweets containing term "allergy" round the year

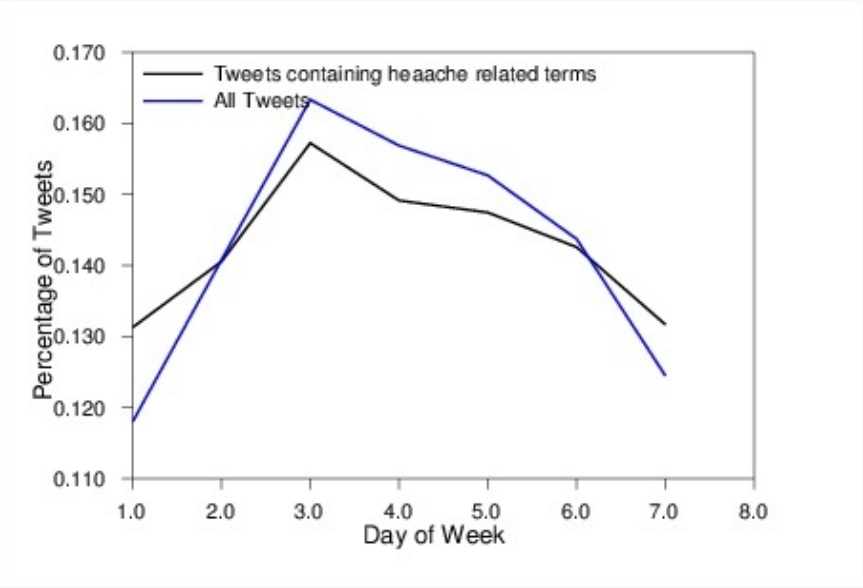
The next ailment patterns we explored are Insomnia and Headache. While we expect the first one to occur during the night hours we do not have any specific hypothesis for the latter one. This basis helped us to provide appropriate time windows for each ailment as shown in each one of the following figures. Figure [13] imprints the distribution of the 7.500 tweets that contain the token “Insomnia” over the seven days of the week starting from Sunday. As we can see insomnia's distribution does not deviate from the distribution that is extracted by the total traffic. For both distributions a clear peak can be observed on Tuesday, followed by a steady decline as time approaches the weekend.



Drawing 13: Distribution of tweets containing term "Insomnia" round the week

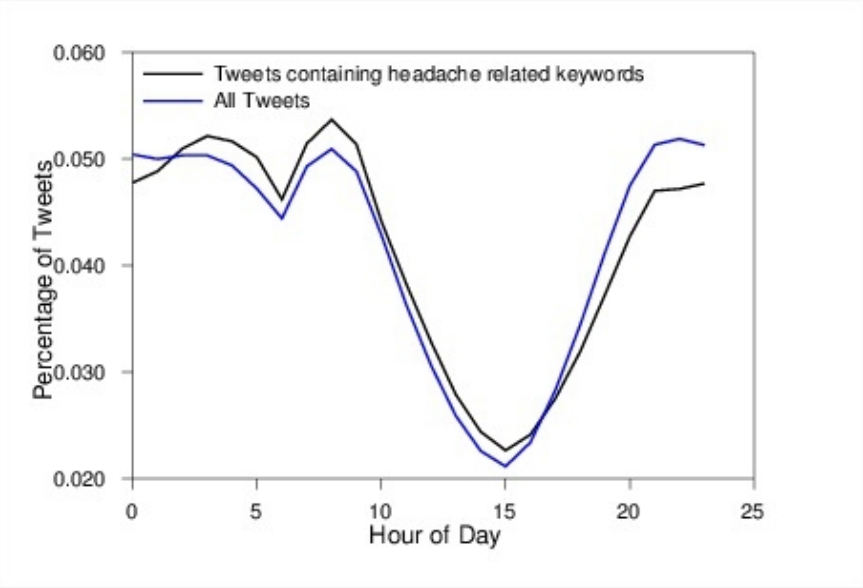
The following figures, [14] and [15] plot the distribution of headache outbreaks on a weekly

and day basis respectively. More specifically, we plotted the distribution of tweets that contained headache related terms over the 7 days of the week and the 24 hours of the day. As we can see, figure's [15] distribution is very similar with the one that is imprinted in figure [2]. Headache outbreaks seem to peak on Tuesday suggesting a correlation between the burden originating from everyday work and the aforementioned well-being aspects.



Drawing 14: Distribution of tweets containing term "headache" round the week

Finally, apart from the aforementioned tendency the most important observation is the striking correlation between the distributions that are extracted from the two corpus subsets. As we can see, distribution over time seems to be independent from the tweet's thematology as tweets related to headache issues follow the same distribution with all other tweets. Moreover, it is clear from the

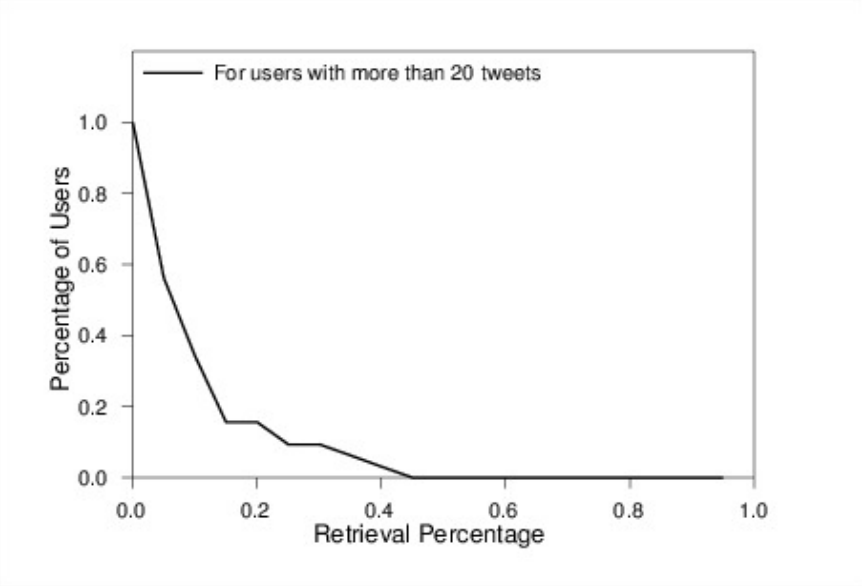


Drawing 15: Distribution of tweets containing term "headache" round the clock

respective plot that posting rate is steady with an exception around the working hours.

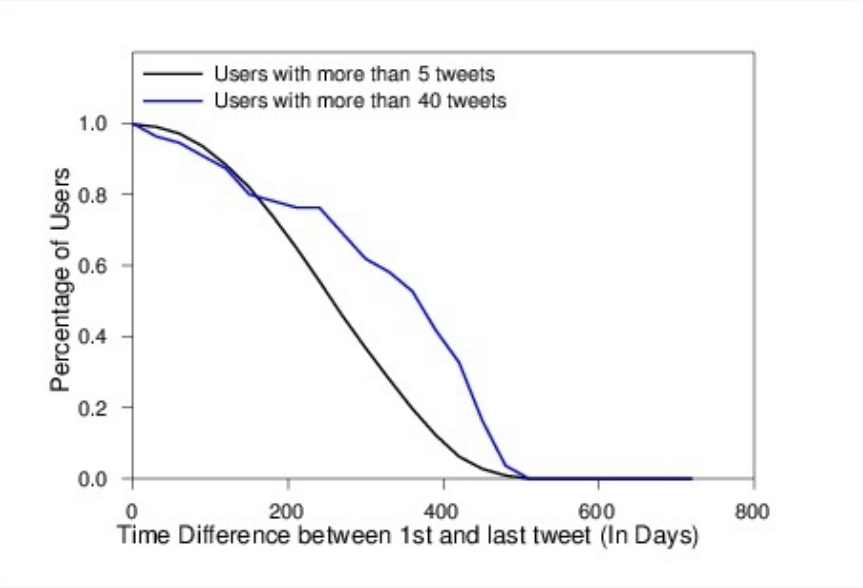
5.1.4 Other Corpus Statistics

Before proceeding with the examination of the remaining corpuses it is important to appose a few remaining statistics. Although the latter ones do not belong semantically to the previously analyzed categories they are crucial for our analysis as they provide in depth insight on further potential exploitation. One of the fundamental aspects that concerned us was the average retrieval percentage per user. This will provide us an insight on the number of remaining tweets that exist for each user. Although no guarantees are provided for their relation to health issues, retrieving those tweets can only be beneficial as those tweets might hold valuable information on other well-being aspects that effect our population's lives. The following figure imprints the retrieval percentage regarding the 'active' users as defined earlier. As we can see, the average retrieval percentage is significantly low leaving us a with a notable potential for further exploration. It is indicative that for only 20% of the people we have a retrieval percentage that exceeds 15%.



Drawing 16: CDF of tweets retrieval percentage for users with more than 20 tweets within the dataset

Another important distribution we present here is related to the total timespan between the first and the last retrieved tweet for each user. This will permit us to identify any time gaps within the corpus extraction process. As figure [17] denotes for the plurality of the users the total timespan corresponds to the actual corpus extraction dates from February 2009 to March 2010.



Drawing 17: CDF of timespan between first and last tweet per user

The last series of statistics we extracted regarding our corpus, are related to the geographical aspect that chaperons our traffic. The latter ones are based on text information each user provides about its current location. Unfortunately, at the same time the pluralism that characterizes the location attribute is its Achilles heel as it requires extensive multilayer address resolving, not even taking into account that many users do not actually provide geographical information. The following table summarizes a subset of the examined geographical aspects of our corpus.

	Total Population	Active Users
Not null location	82.00%	88.00%
Iphone as location	2.10%	3.00%

Table 2: Location statistics

5.2 Examining relations between nutrients and other well-being aspects

Having acquired an initial picture about the trends that exist within the health-related dataset we proceeded with our ultimate goal. As mentioned before the main objective of this thesis is concentrated on the idea of relating different popular nutrients and meals with health or lifestyle aspects. The latter task was tackled using two different approaches, one document-centric which was based on the output of the LDA and another one which was human-centric and was mainly on term matching. In both cases, processes were applied on a subset of our main dataset that was collected over a two-months period. The latter one which was a small fraction of the initial one was a result of a procedure that eliminated any tweet that included no health, lifestyle or nutrition aspect.

5.2.1 Document-centric approach

Fueled by the capabilities and the application-wise versatility that characterizes the Latent Dirichlet Annotation we saw a significant potential into examining relations within different nutrients and other well-being aspects. The main advantage of employing LDA in this direction instead of simply operating on the basis of keywords is concentrated on the fact that LDA model's output is expressed in terms of topics. The latter one permits us to maximize the number of documents that correspond to a certain ailment, nutrient or lifestyle factor and thus minimize as much as possible the statistical error that chaperons such operations. Despite this main feature of the model, there were significant concerns related to the operation's success which were based on the fact that different semantic content may be merged under the same topic as long as co-occurrences among words do exist. Making things even worse, apart from selecting the output's cardinality, a user possesses no real means of directly controlling the output of the model. The only influence on the output is restricted to providing apriori knowledge to the LDA model, in the form of a previously trained model.

In order to tackle with the aforementioned problem, we examined both approaches. While incorporating into the process apriori knowledge did not make any significant difference in terms of output, requesting different numbers of topics did prove successful to some extent. In this direction we initiated the procedure with requesting 50 topics, number which was increased by 50 in each step. The aforementioned tactic that is previously documented in [20] permits us to get a view of the data with a different resolution each time in a process that can be likened with that of a focus mechanism in an analogue photo-camera.

Examining the model's output, although a calculable number of topics were extracted, in the plurality of cases the topics that shape a solid semantic content only constitute a small fraction of the total output. Besides the semantically coherent high frequency topics, the remaining ones were either a mixture of different subjects or repetitions of previously presented topics. Although this aforementioned ascertainment does not attest the absence of other topics in any way, it does though, make clear the void in literature that regards applying topic detection in small size documents that form low frequency topics. In an effort to overcome this partly expected problem we tried to counter-balance the effect of the dominating topics over the less impacting ones. In this context we divided the previously extracted dataset into separate parts using a large number of keywords, resulting in three sub-datasets regarding health, nutritional and lifestyle aspects respectively. Hoping that this step reduces the possibility for a topic to be cloaked by a more impacting one we trained once more an LDA model, one for each derivative starting again from requesting 50 topics. The aforementioned intuitive solution did improve results in a calculable extent, however we believe that its success is highly dependent on the dataset and we though provide no guarantees for the method's success on different sets of documents.

Returning to the actual output, by examining each derivative separately we quickly ascertained that the health-related semantically coherent topics are significantly less than the topics that originate from the remaining two subsets, regarding nutrition and lifestyle respectively. The explanation for the this phenomenon is ambiguous. The latter could be either a result of extensive use of Twitter for marketing purposes as documented recently in [20], or being an immediate positive consequence of good nutrition under the concept that people who are concerned about what they eat, benefit from a better quality of life.

Starting with the health-related derivative, from the six topics that were extracted from the respective sub-dataset the 2 are not relevant to our analysis as judging from the top 20 words, they refer either to Michael Schumacher's injury, or to other injuries that took place during the latest world cup.

Furthermore, the two topics that do seem more coherent, always judging from the list of the 20 most probable words to appear within the topic, refer to problems related to headaches, tiredness, sleep disorders and other general pains that could be related with bad lifestyle choices. Closing the last 2 topics refer to more severe ailments, with different aspects of the latter ones being imprinted in each top-words lists. Having acquired an initial picture of what is discussed regarding health within the generated traffic we proceeded with extracting topics from the lifestyle related sub-dataset. After applying LDA on the latter one requesting once more fifty topics, a smaller subset, including ten of them seemed to be more or less coherent. Contrary to the previous case, only one out of the ten coherent topics refers to an impacting public event and is thus non relevant to our analysis. The remaining 9 solid topics on the other hand cover a wide range of training and fitness concepts with overlap between topics being apparent to some extent. Lastly but equally important, apart from the health and lifestyle aspects we examined the last sub-dataset which corresponds to the nutritional aspects of the collected traffic. Out of the fifty topics that consisted the output of the LDA, ten of them show a coherent content. Moreover, compared to the two previous sub-datasets, the output of the nutrition derivative is characterized by minimal overlap resulting to a wide range semantic content. More specifically, the thematology that constitutes the respective output ranges from coffee drinking to nutrition supplements and from alcohol consumption to workers eating habits, providing us with a rich view of nutritional trends that exist within the society.

Furthermore, having the main objective of our thesis in mind, we proceeded with examining relations among different topics. In order to fulfill this task, after providing a general description for each one of the extracted topics, we concentrated on the document-topic distribution which is included in the model's output. The latter one permitted us to produce a scatterplot visualization of our processed datasets. The desired indication in this case, which would certify that there is an actual relation between different topics is extracted by the trends that characterize the respective scatterplot. More specifically, if for the vast majority of the points that consist a scatterplot, ascending X-axis values correspond to equally ascending Y-axis values, the two participating variables are considered to be positively related.

Unfortunately, after examining more than one hundred different scatterplots we can conclude with some safety that unless a mechanism is found, in order to eliminate topics that are not semantically coherent only minor indications for positive or negative relations can be extracted from the aforementioned document-topic distribution. More specifically, in an important number of the extracted scatterplots, we could observe more than one clear trends. Within them there was usually one which clearly showed the existence of a positive relation as it contained an important number of points, which in this case correspond to tweets, on the diagonal of the respective plot. Moreover, fueled by the the observation that even within ostensibly semantically coherent topics different components exist, we computed the Pearson correlation and the inner product for a small subset of the simingly coherent topics. The following table contains the aforementioned quantitative results. More specifically table [3] includes a percent of the tweets that used any of the n-grams or unigrams within each topic. Although counting tweets that have any of the topic terms may be considered generous, we decided that within the study’s context this measurement provides a relatively simple metric to evaluate the frequency of term usage.

Topic	Most Likely Topic Components	%
6(n)	Coffee,drink,hot,tea,caffeine,cafe,green, cafelife,espresso,latte,instant coffee,coffee addict,starbucks,inspired,iced,coffeegram, latteart,cup,love4coffee	0.01-0.08
43(n)	Fruit,fruits,vitamins,veggies,juice,yogurt,sugar, vegetables,bananas,summer,fresh,nuts,vegan, watermelon,eggs,love,snack,community, strawberries, cheese,	0.003-0.04
9(n)	Beer,drinking,wine,drinks,night,top,glasses,10, big,fact,banana,bottle,find,help,whiskey,cure, friends,drinkers,vodka,hour	0.003-0.05
7(n)	Food,fast,time,15,feel,workers,order,mcdonalds, burgers,avoid,cut,junk,higher,real,ideas,eating, remember,items,worst,unhealthy	0.003-0.09
22(n)	Water,drink,bottle,drinking,body,fact,skin,lemon, cold,sleep,30,levels,metabolism,life,test,will, wake,kit,green,feel	0.003-0.14
43(l)	Workout,gym,fitness,muscle,fit,cardio,fitfam,abs, bodybuilding,fast,lean,motivation,exercise,build, ripped,selfie,building,routine,kim	0.005-0.08
6(l)	Chocolate,bar,bars,energy,protein,12,nutrition, pack,count,clif,chip,dark,peanut,quest,butter, milk,free,double,kind,oz	0.005-0.045
4(l)	Diet,fat,body,loss,weight,food,energy,lose,tips, fitness,healthy,exercise,fast,eat,best,raw, program,simple,ultimate,discover	0.007-0.072
5(h)	Cancer,drugs,breast,patients,support,care, treatment,illness,lung,died,risk,prostate,uk, children,cure,cells,fight,charity,drug,millions	0.004-0.105
22(h)	Headache,time,ive,going,worst,hours,wont,sleep, lot,woke,great,bc,fine,mom,feeling,disease,guys, funny,will,americans	0.002-0.072

Table 3: Comprehensive Dataset Topics Relevant to Health, Nutrition & Lifestyle aspects. The last columnis the range of percentages that corresponds to the tweets that used any of the n-grams or unigrams within each topic.

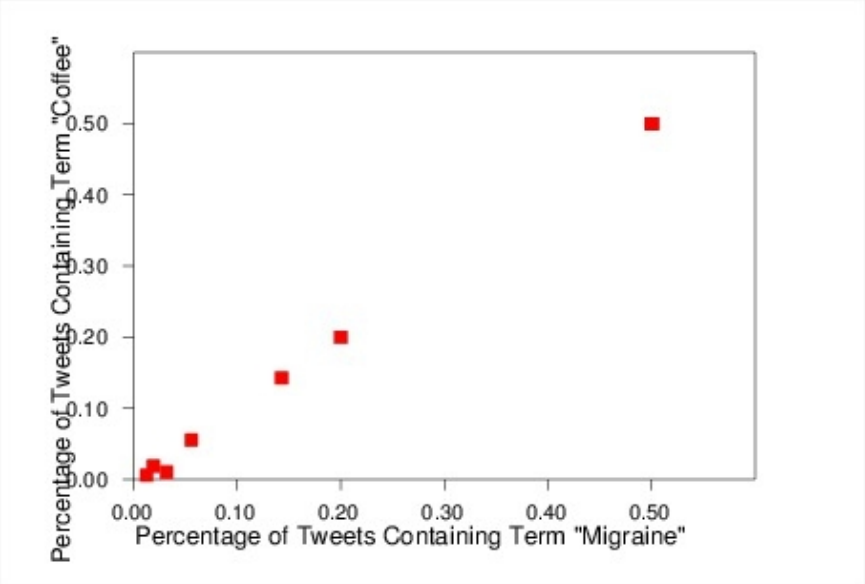
5.2.2 User-centric approach

Realizing the shortcomings of the previously described methodology we decided to deviate from our initial approach. In this case, instead of implementing our approach focusing on documents, we now concentrated on a more user-centric approach. In this context, the points of the desired respective scatterplots did not represent documents anymore. On the other hand, each point now refers to a specific user. Furthermore, the axis now do not represent the correlation with the respective topics. Contrary axis now refer to the percentage of tweets containing a key term for a specific user.

Intuitively, the most demanding task is now related with choosing the appropriate key terms. The importance of the aforementioned procedure is very high as the selection is inextricably related with a successful reveal of the underlying trends. In this context and in an effort to be as objective as possible a set of keywords were carefully selected from the respective lists that contain the most probable words to be generated within each topic, as these topics were extracted in the previously presented steps. While the latter methodology ensures a connection with the previously detected topics the procedure does not suffer from the known LDA's features. As shown in the following figures, for a subset of those words we have indications that a series of publicly known hypothesis that concern nutritional aspects are validated.

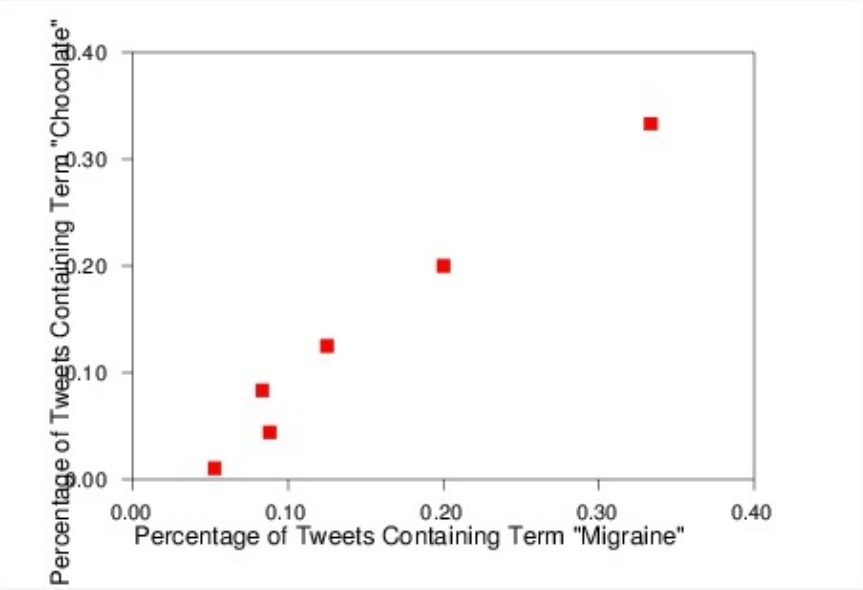
5.2.2.1 Withdrawal syndrome related

The first publicly accepted hypothesis we tried to examine are related to the withdrawal syndrome. The latter one refers to the psychosomatic impact that originates from the deprivation of specific nutrients or other addictive edible consumables like cacao, caffeine, nicotine and sugar.



Drawing 18: Scatterplot, percentages refer to tweets that contain terms "Coffee" and "Migraine" on a user (red dots) basis

Recognizing that coffee has a very solid presence in the LDA's output we examined the relation between the term “coffee” and the term “migraine”. As seen in figure [18], for the users that have at least one tweet containing each term a very clear positive relation is observed. The more term “coffee” participates in one user's tweets the more term “migraine” is encountered.

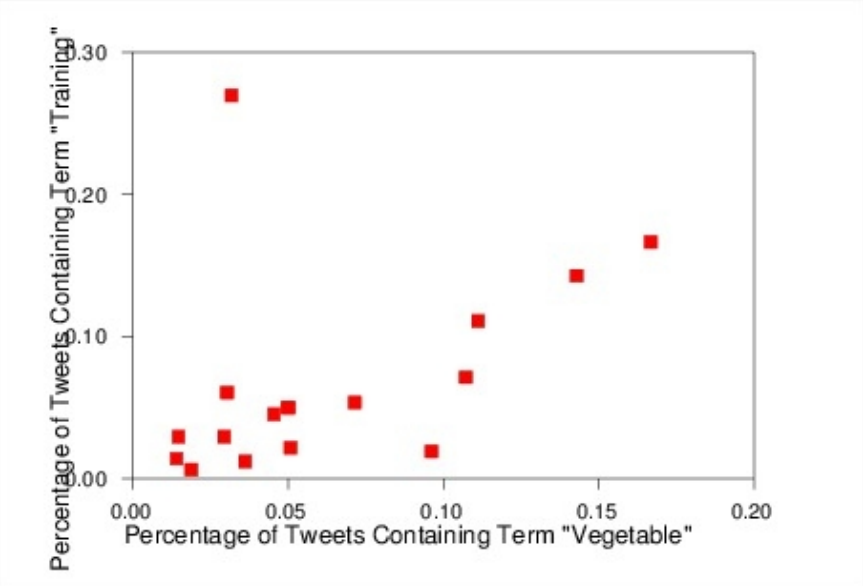


Drawing 19: Scatterplot, percentages refer to tweets that contain terms “Chocolate” and “Migraine” on a user (red dots) basis

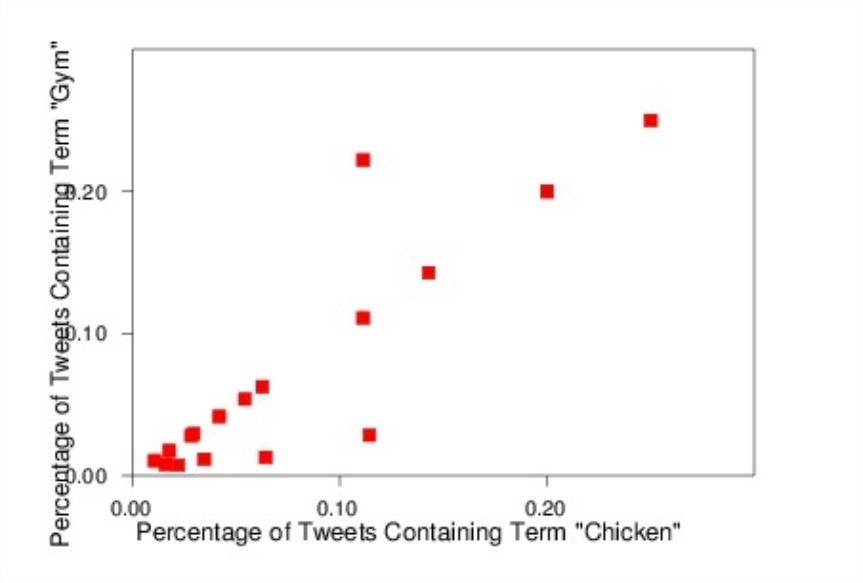
In the same context, another set of terms were examined. As we can see in the respective figure, the term “Chocolate” is positively connected to the term “Migraine” suggesting as previously that users that suffer from migraine may also be suffering from chocolate withdrawal syndrome.

5.2.2.2 Lifestyle related

The second set of hypothesis that we tried to validate are related to “lifestyle”. The latter term condenses a wide range of choices including among others, training, working conditions and other general everyday life aspects.

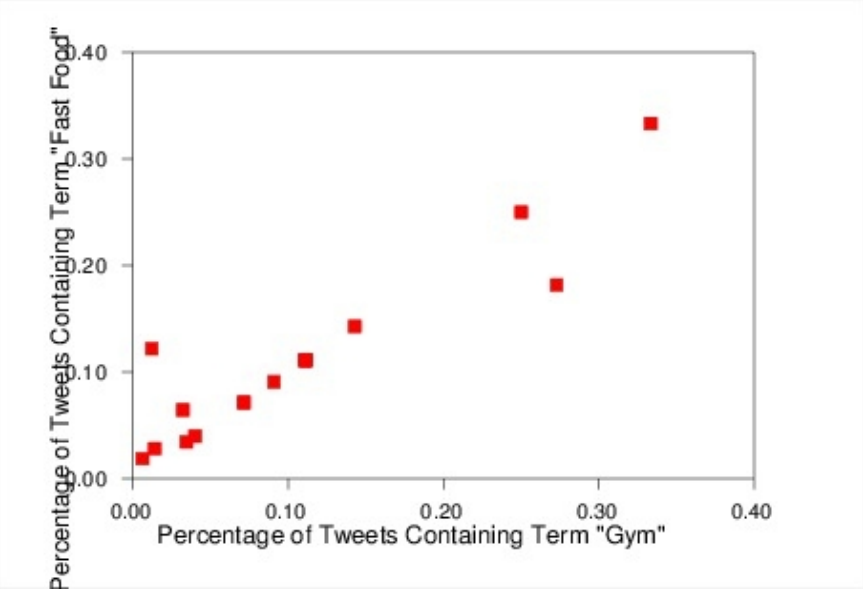


Drawing 20: Scatterplot, percentages refer to tweets that contain terms "Training" and "Vegetable' on a user (red dots) basis



Drawing 21: Scatterplot, percentages refer to tweets that contain terms "Chicken" and "Gym" on a user (red dots) basis

As seen in the respective figures, we tried to correlate fitness aspects with two specific nutrients, chicken and vegetables, which are considered to be ideal for people that train themselves as they are low on sugars and fats but have a high concentration of proteins on the other hand.

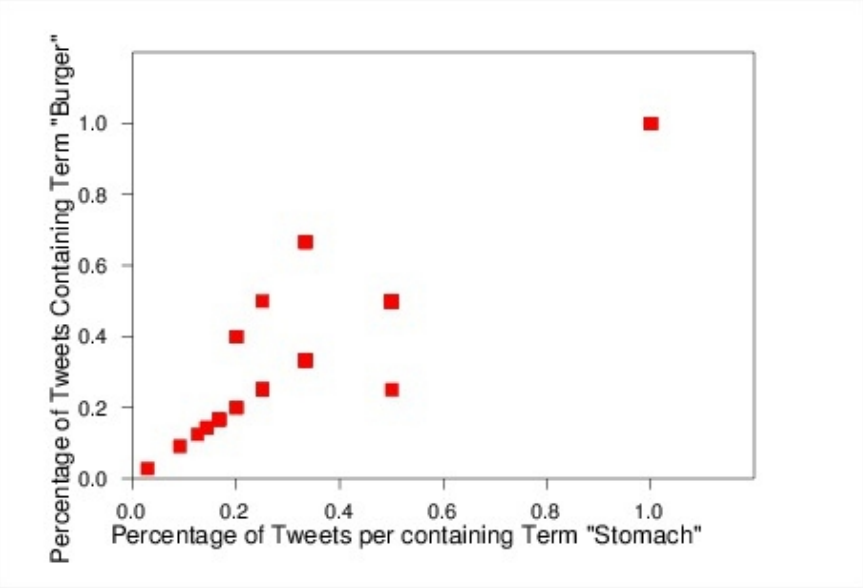


Drawing 22: Scatterplot, percentages refer to tweets that contain terms "Fast food" and "Gym" on a user (red dots) basis

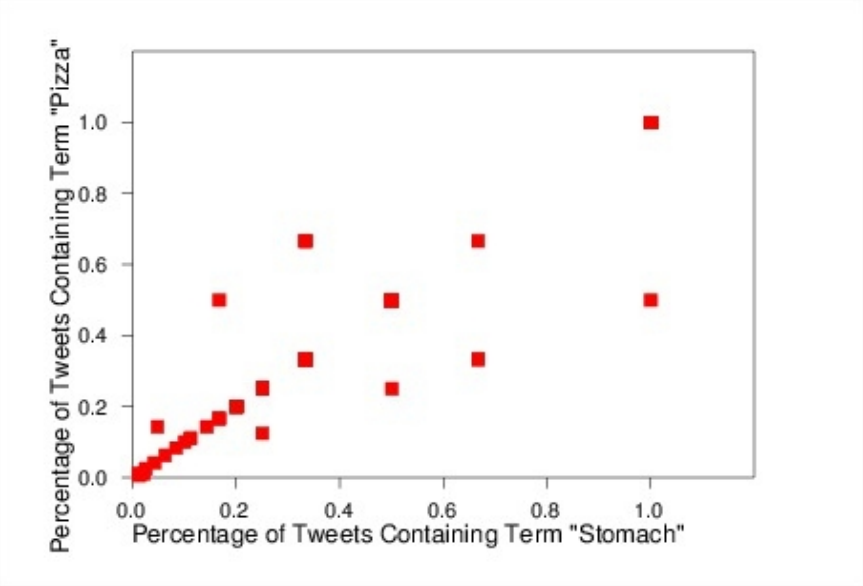
The last pair of words that we examined within this context includes the terms "Fast Food" and "Gym". As seen in the respective figure, both terms are tightly connected through a positive relation suggesting either that people that pay a visit to the gym also tend to eat in fast food restaurants or that people that talk about gyms also talk about the bad impact of junk food in our health.

5.2.2.3 Heavy-eating related

The last category of hypothesis we examined, regard the impact that heavy-eating has on our organism. In this context we plotted the tweets that contain the term "Burger" against the tweets that contain the term "Stomach".



Drawing 23: Scatterplot, percentages refer to tweets that contain terms "Burger" and "Stomach" on a user (red dots) basis



Drawing 24: Scatterplot, percentages refer to tweets that contain terms "Pizza" and "Stomach" on a user (red dots) basis

Closing the current chapter, excluding a relatively small number of exception points, the previous figures show a positive relation between the aforementioned pairs of terms.

6 Thesis Conclusions

For the purposes of this thesis we performed an empirical study on the Twitter OSN which proved to be quite enlightening. During the latter one we collected a large number of micro-blogging documents in order to not only characterize the traffic's synthesis and evolution but also extract some valuable knowledge that derives from the underlying semantically related to health and nutrition activity. By applying a series of state of the art text analysis techniques each one serving a different purpose, we were able to draw some primitive conclusions on a wide range of aspects that collectively compose what is referred as quality of life. Although the aforementioned techniques proved to be more or less effective for uncovering traffic related distributions, a series of shortcomings were observed when trying to extract correlations between nutritional and health variables using the well known Latent Dirichlet Annotation model.

Although literature exhibits a richness of applications that use the aforementioned algorithm, a significant void has been identified in fine tuning this technique for datasets that are originating from micro-blogging social media. The latter ones, as mentioned before, impose new parameters as they consist of an intimidating number of documents and are characterized by a high level of noise. Apart from the encountered shortcomings, one of the most interesting observation regarding the health related dataset was based on the extracted time distributions. Unlike to what was expected, a series of ailments which are sometimes believed to be nutrition dependent seem to be affected by the psychological burden that is imposed by the weekly working routine. More specifically it is observed that most of the health related references and complains are encountered around the Tuesday and Wednesday and start to fade out as time approaches Weekend.

Based on the analysis of the topic-tweets distribution on one hand and of the distributions of a set of hand-picked key terms within the users' activity on the other hand, interesting positive relations between specific ailments and nutrients were found. Among the latter ones, we distinguished those that refer to the withdrawal syndrome including that of caffeine and chocolate. In this context, analysis results contain strong indications that the lack of the latter ones causes migraine. Besides withdrawal syndrome, other nutrition patterns were investigated including nutrition followed by people that tend to train themselves and nutrients that are suspected to be digested with difficulty. While investigating the first aspects, less strong positive relations were found correlating chicken and vegetables consumption to training. Lastly but equally interestingly, indications suggesting that pizza and burger eating imposes stomach problems were observed.

All the above show that there is still plenty of room for improvement of existing solutions and for the derivation of new ones that would be able to capture the dynamic evolution that characterizes graphs from various new and emerging application domains like social networks and other types of web 2.0 applications.

Bibliography

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media" Business Horizons, 2010.

[2] B. Buter, N. Dijkshoorn, D. Modolo, Q. Nguyen, S. van Noort, B. van de Poel, A. Ali, and A. Salah1. Exploratory visualization and analysis of a social network for arts: The case of deviantart. Journal of Convergence Volume, 2(1), 2011.

[3] Hila Becker, Mor Naaman, Luis Gravano: Learning Similarity Metrics for Event Identification in Social Media. WSDM 2010 .

[4] Sasa Petrovic, Miles Osborne, Victor Lavrenko: Streaming First Story Detection with Application to Twitter. HLT 2010 .

[5] Vance K, Howe W, Dellavalle RP. Social Internet sites as a source of public health information. Dermatologic Clinics, 2009 Apr;27(2):133-6. PMID: 19254656.

[6] Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen, Intern Med. 2011 Mar;26(3):287-92. doi: 10.1007/s11606-010-1526-3. Epub 2010 Oct 13.

[7] Fernandez-Luque, L.; Karlsen, R.; and Bonander, J. 2011. Review of extracting information from the social web for health personalization. Journal of Medical Internet Research 13(1).

[8] Jain, S. H. 2009. Practicing medicine in the age of facebook. New England Journal of Medicine 361(7):649–651.

[9] Hawn, C. 2009. Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. Health Affairs 28(2):361–368.

[10] Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014.

[11] Carneiro, H., and Mylonakis, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–64.

[12] Pelat, C.; Turbelin, C.; Bar-Hen, A.; Flahault, A.; and Valleron, A.-J. 2009. More diseases tracked by using google trends. Emerg Infect Dis 15(8):1327–1328.

[13] Chew, C., and Eysenbach, G. 2010. Pandemics in the age of twitter: Content analysis of tweets using the 2009 h1n1 outbreak. PloS ONE 5(11):e14118.

[14] Lampos, V., and Cristianini, N. 2010. Tracking the flu pandemic by monitoring the social web. In IAPR 2nd Workshop on Cognitive Information Processing (CIP 2010).

[15] Culotta, A. 2010a. Detecting influenza epidemics by analyzing twitter messages. arXiv:1007.4748v1 [cs.IR].

[16] Culotta, A. 2010b. Towards detecting influenza epidemics by analyzing twitter messages. In KDD Workshop on Social Media Analytics.

[17] Quincey, E., and Kostkova, P. 2010. Early warning and outbreak detection using social networking websites: The potential of twitter. In Electronic Healthcare. Springer Berlin Heidelberg.

[18] Ritterman, J.; Osborne, M.; and Klein, E. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In Workshop on Mining Social Media.

[19] Michael J Paul, Mark Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. ICWSM 2011.

[20] Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. In Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction (SBP'11), John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai (Eds.). Springer-Verlag, Berlin, Heidelberg, 18-25.

[21] https://dev.twitter.com/docs/api

[22] https://dev.twitter.com/docs/api