Smoking Cessation Causes: Contrasting Evidence from Social Media and Online Forums

Foaad Farooghian[1] and Mourad Oussalah[2]

[1] iHR, Faraday Wharf, Innovation Birmingham Campus Holt Street, Birmingham Science Park Aston, Birmingham, B7 4BB (UK)
[2] University of Oulu, Centre for Ubiquitous Computing, Computer Science, PO Box 4500, 90014 Oulu (Finland)
Mourad.Oussalah@ee.oulu.fi

Abstract. An automated-based system has been developed in order to gather stories from iCanQuit. Similarly, using Twitter Streaming API, geolocated tweets within UK region have been collected during a three months period, and those related to smoking cessation, according to the semantic text matching, are extracted and stored in a database. An automated classifier has been employed to identify the four most dominant categories in Health field for each document of blogs or Twitter dataset. A total of 880 stories from iCanQuit and 22155 relevant tweets have been collected and indexed in SQLite database. The automated classifier highlighted four categories: Weight Loss, Mental Health, Addiction, Support Group. The analysis surprisingly reveals that both blogs and Twitter datasets agree that the dominant source of smoking cessation is related to weight loss (shape body appearance), or *body-look*, while the support group, which includes any clinician supports, plays little impact on the smokers' quit motivation, a result that maybe precious for health authorities.

Keywords: blogs, Twitter, health, smoking

1 Introduction

Tobacco remains one of the prime causes of death worldwide causing more than 5 million deaths annually in 2012 and expected to reach 10 million by 2030 according to world health organization [5] in additional to tremendous economical costs. This motivates the growing smoking cessation initiative programs. In the age of E-generation, Web-based smoking cessation programs have been found to raise smokers' consciousness about quitting and encourage them to take necessary actions [8]. Indeed, online social networks offered a new way to interact with smokers and influence their behaviour where social network intervention may work through multiple mechanisms, including social support, information transfer, social influence, modelling, and the transmission of social norms [7]. Several individuals, especially teenagers, are more vulnerable into following their mates' routes and reasoning rather than bothering to discuss

detailed personal experiences with clinical experts. Besides, it is commonly acknowledged that many individuals feel more comfortable when interacting via text or online messages instead of physical one-to-one meetings. For instance, several mobile applications have been promoted for this purpose (e.g., I QUIT, Quint O Meter, Quit With Me, UbQUITous), that would ease the interaction of smokers willing to quit [9], although their compliance with clinician guidelines is debatable. Therefore, there is a recognized need for research on the use of social media to promote health behaviors and social support [14]. Platforms like QuitNet (`www.quitnet.com`) [2] have attracted millions of smokers seeking help, share of stories, alternative remedies, among others, mainly for the purpose of smoking cessation. Similarly, iCanQuit (`www.icanquit.com.au`), is another popular online service provided by Australia cancer institute NSW since 2010 [11]. It is characterized by its easy access and well-structured documents, with many classes as opposed to QuitNet. The platform also offers the possibility of retrieving past data from users and clinicians. Finally, the expansion of the social media tools, e.g., Twitter, Facebook, Flickr provides the analyst with a huge amount of data related to daily behaviour of the individuals /smokers, which can be efficiently employed for behavioural analysis of the smoker (s), and thereby, prescribe appropriate remedies accordingly. Therefore, the analysis of blogs and other social network data related to smoking cessation opens a new door for researchers to understand the reasons that motivate people to smoke. This also enables scientists and/or clinicians to identify new symptoms or effects during smoking cessation which could result in possibly more efficient treatments. This paper presents a blog and a social media related study that focused on symptoms and/or origin of smoking behaviour/cessation by investigating data issued from both Twitter Streaming API and the specialized smoking quit platform IcanQuit. In both cases, parsers and crawler were employed to collect the messages. A textual analysis of the messages is carried out in order to identify the cause (s) of the smoking behaviour using a classification like strategy. Comparison of the classification results in both sources (Twitter and blogs) is carried out in order to identify key milestones.

2 Methods

2.1 Data collection

Two sources of information have been employed for data collection and analysis. The first one consists of dataset gathered from IcanQuit blogs. Interestingly, this website has already some predisposition of smoking quiet stories with respect to a set of predefined categories: health, money, fitness, family and others. In contrast to data that can be retrieved from QuitNet or other blog sources, the stories in IcanQuit are well-structured texts with a title and sometimes (sub) sections. A total of 880 stories have been collected. Fig. 1 provides an example of the website configuration as well as an example of database outputted by iCanQuit query related cessation story where the relevant links are therefore stored in a SQLite database. The second source of information uses Twitter Streaming

API [3] to collect geolocated tweets in UK region for a specified time interval. Next, the collected Twitter database has been filtered using smoking and quieting related terms in the text message field. More specifically, Twitter dataset has been collected over the period August - November 2014, and then filtered such that any relevant tweet contains at least one "smoking" related terms and one "quiet" related terms. The former includes terms like: smok, cigarette, cig, tobacco, cannabis, pipe, shisha, waterpipe, hookah, and weed (inspiring from urban dictionary). While quiet related terms comprise both "quiet" equivalent terms and commonly employed smoking cessation products, e.g., quiet, cessation, cess, stop, nicotin, patch, lozeng, chantix, counselling, therapy. A total of 3245567 tweets have been collected among which 22155 have been found to fit the smoking cessation context using the above methodology. These tweets are then stored in a SQLite database.

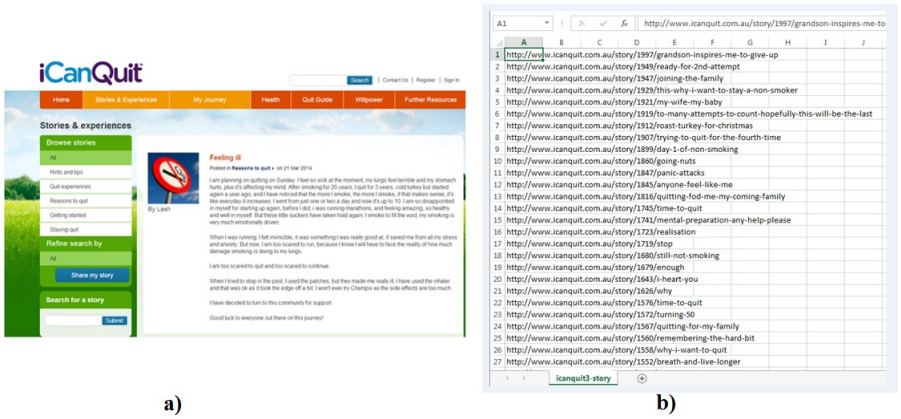

Fig. 1. Example of a) iCanQuit configuration and b) output of crawler of iCanQuit.

Unlike structured story documents from iCanQuit platform, Twitter dataset, due to the size restriction (140 characters at most), are dominantly noisy with a lot of slang words, abbreviations, spam and links, which require a special methodology to capture relevant information. Inspired from our previous work [12], special consideration has been given to text pre-processing of textual Tweet messages. Especially, Appache Nutch Crawler [1] was employed in order to crawl the link extracted from iCanQuit website. The extracted text, usually constituted of user's experience and story, is parsed using open source Stanford Parser and then indexed using Apache Lucene [10] in order to benefit from its highly scalable implementations and advanced search capabilities. The created index files are stored in SQL like database that eases the compatibility with other software resources.

2.2 Evidence classification

The text tokens outputted from previous stage using either iCanQuit or Twitter are classified according to a set of predefined classes. For this purpose, the UClassier API [4] was employed. This implements an improved semi-supervised nave Bayes classifier with an extensive training corpus issued from yahoo open directory project (`http://www.dmoz.org`). We restricted to health topics of the directory because of its relevance to smoking cessation. Especially, we confine our analysis to the four classes which exhibit high score for most of the input dataset. This corresponds to: Mental Health, Weight Loss, Addiction and Support Group. Especially, a given document is deemed to belong to a specific category if the associated classification score according to UClassifier is the highest among other categories and is greater than some pre-defined threshold (a 10%. threshold is found to give satisfactory results). The use of threshold is motivated by the existence of documents related to smoking cessation but they do not contain any argument that would allow the system or, even any expert who reads the document, to identify any possible causes for smoking cessation. On the other hand, the analysis of dataset involves two main phases. The first phase examines the relationship between different classes based on the Uclassifier classification by retrieving the score attached to each document. Next, the score of each class is recorded, and the correlation between each pair of the four aforementioned classes is examined.

2.3 Data Correlation Analysis

A final stage consists in a statistical analysis of evidence issued from the classifiers. For this purpose, the correlation matrix among the various classes is calculated for both iCanQuit and Twitter datasets. On the other hand, the evaluation of the extent to which the two datasets support the same evidence is also quantified using statistical testing. More specifically, Cramer's phi correlation test [13] is employed to evaluate the (global) correlation between Twitter and iCanQuit dataset, while scatter graphs and Pearson's product correlation coefficient is employed to quantify the correlation among any pair of categories using either Twitter or iCanQuit dataset. The evaluation of the result of the automated classification against manually labelled classification is performed using groups of random selection of dataset and accuracy classification metric.

3 Results and Discussions

Given that not all collected dataset contain enough clues to enable the system to generate the cause of the smoking cessation, it is worth pointing out the proportion of data where such evidence is occurring. For this purpose, Table 1 (see first column) shows the number of blogs and tweets falling in each of the four categories. The results in terms of total number of blogs and tweets in the four categories indicate that 53.7% $((273 + 86 + 72 + 42)/880)$ of retrieved blogs

and only 17% ((1730 + 903 + 906 + 234)/22155) of retrieved tweets are classified to one of the four categories. Trivially, as expected, larger this to occur more often in Twitter dataset because of wording size. Assuming the outcomes of blogs and Twitter dataset lie on two random variables, Spearman's rank correlation coefficient can be used to quantify their correlations. In this respect, a very strong correlation is observed $\rho = 0.93$ with $p < 0.07$. Table 1 indicates a strong dominance of Weight Loss or, by abuse, *body look* as the main catalyst for change of smoking behaviour including smoking cessation. This is also a well-known fact in tobacco community as the effect of nicotine increases metabolic rate in the human body, which, in turn, suppresses appetite and yields weight loss. While smoking cessation is likely to induce a gain of weight [6]. In order to investigate the relationships between pairs of categories while accounting for the sensitivity of the scoring function of the classifier, we created nine equal partition of the (normalized unit interval) of the scoring function in [0.1 1], and record the number of blogs (resp. Tweets) whole classification score of the underlying category fails in the given subdivision. This also allows us to use the Pearson product moment correlation coefficient to evaluate the correlation between any pair of categories. In this course, Table 1 also records the correlation matrix for both blog and Twitter dataset (the latter is in bold). The result indicates a moderate correlation between categories Mental Health and Weight Loss as well as between Support Group and Weight Loss while the correlation among other categories is quite weak to negligible. In both scenarios, this demonstrates the prevalence of body shape argument (weight loss) for clinical analysis as well as the importance to ease the effects of mental health difficulty.

Category # blogs vs #Tweets	Weight Loss	Addictions	Support Groups	Mental Health
Weight Loss 273 vs **1730**	1			
Addictions 72 vs **906**	−0.142 −0.123	1		
Support Groups 42 vs **243**	−0.207 − **0.245**	0.0255 **−0.005**	1	
Mental Health 86 vs **903**	−0.154 −0.183	−0.078 −0.032	−0.099 −0.121	1

Table 1. Correlation matrix between categories using blog dataset and Twitter dataset (shown in bold) ($p < 0.001$).

On the other hand, it is also worth pointing out that there is a substantial amount of dataset that cannot be classified to any of the aforementioned categories, where about 46% of blogs and 82% of Twitter messages have not been classified. This is mainly motivated by the nature of the dataset where many of the messages and stories are too short or contain no useful information that

would allow any classification system to yield a specific category. For instance, messages "Yes I tried to quit cig", "this is a good story of smoking cessation attempt" or "how was your day after first quit?" provide little information for any external user to output any tangible conclusion regarding the cause of cessation. Similarly, there are a large number of tweets which are mainly generated by commercial organizations to promote some specific smoking cessation products, e-cigarette, nicotine brands, etc. Some stories are also found to be related to the subject through their title but the story itself is very short and can be restricted to a web link only. Finally, in order to compare the accuracy of the global category classification of Table 1 against manual check constituted by an independent expert, we selected a set of random samples from the dataset and computed the accuracy in the following way. For each category, we decomposed the associated dataset into three (almost) equal groups. For instance, the 273 blogs corresponding to Weight Loss category are split into three equal groups of 91 blogs each. For each group, we randomly selected 10 documents, which are manually checked, and then its accuracy is computed. This process of randomly selecting 10 documents from each group, and then calculating the associated accuracy, is repeated for each category. Table 2 summarizes the accuracy of these subgroups when using blogs and Twitter dataset. Results highlighted in Table 2 demonstrate a good accuracy of the automated classification system when compared to a manual expert-based classification. We also acknowledge a slightly decreasing accuracy in case of Twitter dataset. This can be explained by the difficulty of the tweet message to be classified in one of the specified category because of the size restriction which renders any manual classification rather a difficult task and sometimes very subjective.

	Blogs dataset			Twitter dataset		
	G.1	G.2	G.3	G.1	G.2	G.3
Weight Loss	93%	96%	95%	87%	81%	83%
Mental Health	95%	94%	94%	78%	82%	87%
Addiction	92%	98%	97%	84%	85%	80%
Support Group	96%	99%	95%	77%	76%	83%

Table 2. Accuracy of the category classification.

Acknowledgment

This work is supported by University of Birmingham School of Electronics, Electrical and Computer Engineering as well as EPSRC GAP Project, which are grateful for their financial help in conducting this research, when the first author was with Birmingham University.

References

1. Apache Nutch Crawler. http://nutch.apache.org/. Accessed: April 2014.
2. Quitnet, HELP FAQs. http://www.quitnet.com/help/helpfaq.jtml?SubjectID=925. Accessed: April 9, 2014.
3. Twitter Streaming API. https://dev.twitter.com/streaming/overview. Accessed: April 2014.
4. UClassifier API. http://www.uclassify.com/About.aspx. Accessed: April 2014.
5. WHO, Tobacco fact sheet N339. http://www.who.int/mediacentre/factsheets/fs339/en/index.html. Accessed: August 2014.
6. B. Borrelli and R. Mermelstein. The role of weight concern and self-efficacy in smoking cessation and weight gain among smokers in a clinic-based cessation program. *Addictive Behaviors*, 23(5):609 – 622, 1998.
7. N. K. Cobb, A. L. Graham, and D. B. Abrams. Social network structure of a large online community for smoking cessation. *American Journal of Public Health*, 100(7):1282–1289, 07 2010.
8. C. Escoffery, L. McCormick, and K. Bateman. Development and process evaluation of a web-based smoking cessation program for college smokers: innovative tool for education. *Patient Education and Counseling*, 53(2):217–225, 2004.
9. A. M. Jacobs, O. C. Cobb, L. Abroms, and L. A. Graham. Facebook apps for smoking cessation: A review of content and adherence to evidence-based guidelines. *J Med Internet Res*, 16(9):e205, Sep 2014.
10. M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA, 2010.
11. K. M. McElwaine, M. Freund, E. M. Campbell, C. Slattery, P. M. Wye, C. Lecathelinais, K. M. Bartlem, K. E. Gillham, and J. H. Wiggers. Clinician assessment, advice and referral for multiple health risk behaviors: Prevalence and predictors of delivery by primary health care nurses and allied health professionals. *Patient Education and Counseling*, 94(2):193–201, 2014.
12. M. Oussalah, F. Bhat, K. Challis, and T. Schnier. A software architecture for twitter collection, search and geolocation services. *Knowl.-Based Syst.*, 37:105–120, 2013.
13. D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 4 edition, 2007.
14. J. L. Westmaas, J. Bontemps-Jones, and J. E. Bauer. Social support in smoking cessation: Reconciling theory and evidence. *Nicotine & Tobacco Research*, 12(7):695, 2010.