How Joe and Jane Tweet about Their Health: Mining for Personal Health Information on Twitter

Marina Sokolova
CHEO Research Institute
University of Ottawa
sokolova@uottawa.ca

Stan Matwin
Dalhousie University
University of Ottawa
stan@cs.dal.ca

Yasser Jafer
University of Ottawa
yjafe089@uottawa.ca

David Schramm
CHEO and University of Ottawa
dschramm@toh.on.ca

Abstract

With 19%–28% of Internet users participating in online health discussions, it became imperative to be able to detect and analyze posted personal health information (PHI). In this work we introduce two semantic-based methods for mining PHI on social networks which will warn the users about potential privacy breaches. One method uses WordNet as a source of health-related knowledge, another - an ontology of personal relations. We use Twitter data to empirically evaluate our methods. We also apply Machine Learning to demonstrate advantages of our extraction procedure when tweets containing PHI have to be automatically identified among other tweets.

Keywords: Text mining, Twitter, Personal Health Information, Machine Learning

1 Introduction

Online networking websites *Facebook, Twitter, PatientsLikeMe* became popular communication hubs connecting millions of individuals. In casually written messages (posts, tweets, updates), people discuss life experience (`i plan to stay home and watch christmas movies while I get better`) and comment on various events (`The discovery was made via CAT scans`) [1] amongst others.

While posting about personal health, a user reveals details that in pre-social network era were usually discussed during visits to a health care provider or in a family setting. This detailed health description is called Personal Health Information (PHI) (Hersh, 2009). Posted online PHI is used in several practical applications: formulating Web policies, including privacy and confidentiality concerns or information leak prevention (Ghazinour et al, 2013a), understanding population response on health care policies (vaccination, immunization)(Chew and Eysenbach, 2010), and an early detection of adverse health-related events (Lampos and Christianini, 2010).

Recent studies of 11,000 posts on a social network showed deficiencies of traditional electronic sources of medical information in the task PHI detection (Ghazinour et al, 2013b).

Our current work aims to show that it is possible to considerably improve accuracy of PHI extraction from social networks. Our approach uses the PHI ontology presented in (Sokolova and Schramm, 2011). The ontology's structure and terms reflect on patient communications in health care setting. In this paper, we present two semantic-based enhancements of the ontology and apply them to extract PHI in Twitter. One enhancement uses WordNet as a source of health-related knowledge, another – an ontology of personal relations.

We use manual analysis to demonstrate that incorporating semantic information significantly improves *Precision* and *Fscore* of the PHI text retrieval. We also apply Machine Learning methods to show the advantage of our approach in automated detection of PHI. Partial preliminary results of this work had been reported in (Sokolova et al, 2012).

2 Vocabulary Resources for PHI detection

It has been estimated that 19% – 28% of all Internet users participate in medical online forums, health-focused groups and communities and visit health-dedicated web sites (Baliccon and Paganelli, 2011),(Renahy, 2008). People share their health-related worries and medical conditions.

[1] All messages have authentical spelling and content.

Person		Diseases and Related Problems		Health Care System	
Anatomical parts	head, kidneys	Diseases	arthritis, depression	Providers	dentist, surgeon
Physiological functioning	insomnia, pregnancy	Symptoms	fever, pain	Setting	ambulance, hospital

Table 1: Examples of categories and terms of the PHI ontology

Automated text analysis often uses only a key word search to find PHI in the user posts. For example, in relation to the H1N1 pandemic in 2008–2009, occurrences of the PHI textual markers (fever, temperature, sore throat, flu) were traced in several geographic areas (Lampos and Christianini, 2010). The extracted tweets, however, were not analyzed if they indeed contain PHI, and all the retrieved messages were considered equally important.

A simple example illustrates limitations of the key word search. The following messages are extracted with a keyword flu: (you are funny comparing the Iphone to a flu shot lol nice) and (I am trying to recover from my turn with the flu).

Whereas the latter message is relevant to personal health, the former is not, but both were counted towards the flu symptoms.

The use of specialized resources of health-related terms can focus the analysis by refraining from the extraction of text irrelevant to PHI. In (Ghazinour et al, 2013b), the authors applied semantic analysis and domain knowledge, to find MedDRA and SNOMED terms related to personal health.

Below we compare the effectiveness of both resources with the PHI ontology (Sokolova and Schramm, 2011). The ontology contains a four–level hierarchy of concept categories corresponding to health discussions by the general public. The categories reference to anatomical parts and physiological functioning of body, diseases and symptoms, and the health care system. Extensive clinical experience of one of the authors was applied to empirically adapt the taxonomies to patients description of their health. As a result, the ontology contained 500 terms commonly used by patients in clinical setting. Table 1 lists two upper-level categories and examples of terms.

It should be emphasized that the presence of one or more health ontology term(s) does not necessarily guarantee that this tweet refers to personal health. In well Im keeping my eye on you just so you know, the word eye indicates ”anatomical body part” but the message does not refer to personal health. Therefore, manual screening of the extracted messages is required in order to remove irrelevant messages.

We worked with the Twitter data from the Content analysis of Web 2.0 workshop [2]. The data was organized as threads, i.e. consecutive tweets posted by users. Only conversational tweets were present; spam, ads, organizational and promotional tweets were discarded. In this work, we use the tweet content, but not the meta characteristics (e.g., time and geo-locations of tweets).

We manually analyzed usefulness of health terms in extraction of tweets containing PHI. The original Twitter set has been organized in threads; hence, we used this unit in the selection step. To decrease an impact of a possible selection bias, we ran five rounds of random thread selection. Each round selected 200 threads. For each selected set, we extracted tweets with the health terms. 3017 tweets were extracted in total, from those 889 tweets contained PHI. Based on the manual analysis, the performance was evaluated by

$$Coverage = \frac{|Extracted\ texts|}{|Texts\ in\ corpora|} \tag{1}$$

$$Precision = \frac{|Extracted\ texts\ with\ PHI|}{|Extracted\ texts\ |} \tag{2}$$

$$Recall = \frac{|Extracted\ texts\ with\ PHI|}{|Texts\ with\ PHI|} \tag{3}$$

$$F - score = \frac{2PrecisionRecall}{Precision\ + Recall} \tag{4}$$

The extraction results were consistent across all the five subsets and significantly more accurate

[2] http://caw2.barcelonamedia.org/node/7

Tools	# of terms	Texts in corpora	Extracted texts	Extracted texts with PHI	*Coverage*	*Precision*	*F-score*
MedDRA-PHI	8561	11000	744	86	0.068	0.12	0.21
SNOMED-PHI	44802	11000	673	108	0.061	0.16	0.28
PHI ontology	500	36315	3017	889	**0.083**	**0.30**	**0.46**

Table 2: PHI extraction using MedDRA-PHI, SNOMED-PHI, and the PHI ontology terms. *Recall* = 1.00 for the three sources. MedDRA and SNOMED results are adapted from (Ghazinour et al, 2013b)

PHI ontology	Performance improvement		
vs	*Coverage*	*Precision*	*F-score*
MedDRA-PHI	122%	250%	220%
SNOMED-PHI	136%	188%	164%

Table 3: Advantage of the use of the PHI ontology in extraction of PHI texts

Data	*Precision*	*F-score*
All PHI tweets	0.30	0.46
PHI tweets with the PO	**0.4**1	**0.58**
PHI tweets sans the PO	0.25	0.40

Table 4: Impact of the PO terms in extraction of PHI texts.

than those of MedDRA-PHI and SNOMED-PHI. Table 2 presents the results of the extraction, Table 3 exemplifies benefits of the PHI ontology over MedDRA-PHI and SNOMED-PHI in extraction of texts containing PHI.

Manual analysis of the extracted tweets revealed that most of tweets that do not reveal PHI were extracted with the PHI ontology terms from the Body and Organs categories. Among them, head, hand, heart were the top contributors to extraction of non-relevant messages (e.g., back to work, lolo get outta my head).

3 Semantic Enhancement of the PHI Extraction

In the current work we wanted to improve *Precision* of the extraction, without jeopardizing *Recall*, and reduce dependance on a manual analysis. We decided to reinforce the lexicon-based search with semantic enhancement. We used enhancement specific to PHI disclosure: a) a set of personal references organized as ontology of personal terms (Section 3.1), b) health terms' semantic information provided by WordNet (Section 3.2). We used *Precision(Pr)*, *Recall(R)*, *Fscore(F)* to evaluate the performance.

3.1 Ontology of Personal Terms

We observed that in messages discussing personal health, a user often directly refers to the person whose information is disclosed. This could be the user itself (e.g. appointment at the plastic surgeon today for my scar from the accident) or relatives (e.g. my oldest had his th bday today & he had the stomach flu).

We marked such references and then organized them in Ontology of Personal Terms (PO). At this point, the ontology includes terms representing the relationship between the user and family members. The terms were divided into four lexical categories, namely, Subjects, (e.g. I, he, she), Possessive Determiners (e.g. my, his, her), Relatives (e.g. son, daughter, parents), and verbs of belonging (e.g. has, have, was).

We expected a higher accuracy of detection and extraction of health information related to an individual when the health ontology is enhanced with the personal ontology. We started with incorporation of the PO terms into the tweet retrieval. On this step, we were looking for the impact of personal terms on retrieval of tweets with PHI. We grouped all the tweets retrieved with PHI terms into two sets: with explicit personal reference (i.e., with PO terms), such as (I am trying to recover from my turn with the flu), and without explicit personal reference (i.e., no PO terms), such as (PSA tylenol cough & sore throat has more cough suppresant than all overthecounter cough syrups).

We manually analyzed how PO terms contribute to the accuracy of extraction of tweets with PHI. Presence of the PO terms in PHI tweets increased *Precision* by 64%, *F-score* – by 45 % (Table 4).

terms	# of synsets
`Allergy, Hospital`	1
`Anxiety, Fever`	2
`Dizzy, Emergency`	3
`Sore, Panic`	4
`Tooth, Itching`	5
`Diet, Stomach`	6
`Infection, Pain`	7
`Hurt, Stress`	≥ 8

Table 5: Examples of the PHI terms and the number of their synsets.

3.2 Semantic Information from WordNet

WordNet[3] groups words in sets of cognitive synonyms (i.e., synsets), builds super-subordinate relations of the synsets, differentiates between common nouns and specific instances, etc. Each term has a number of corresponding synsets; the synsets are ordered from the most common to the least common. For example, the word `fever` has the representation:

- S: (n) fever, febrility, febricity, pyrexia, feverishness (a rise in the temperature of the body; frequently a symptom of infection);
- S: (n) fever (intense nervous anticipation) (in a fever of resentment).

The representation shows that `fever` more often signifies a rise in a body temperature than a nervous anticipation.

The number of synsets is a strong indicator of the number of different senses of the word (i.e. ambiguity). For health terms, a lesser number of synsets show a stronger correspondence of the term to personal health information. Table 5 lists examples of the health terms and the number of their synsets.

The rank of the health-related synset among the all synsets of the term is another strong indicator of the usefulness of the term in the given context. For example, `fever` has the rank 1 as its health-related synset is 1. Preliminary observations showed that 1st rank of the term's health synset is a strong indicator of the term relevance to personal health information.

3.3 Evaluation of Semantic Enhancement

To assess how accurate health terms are in the recognition of the tweets with PHI, we looked at the number of synsets and the health-related rank of the terms. We then manually analyzed tweets extracted with health terms and subdivided them into those with PHI and others. To follow the impact of the number of synsets and the health-realted rank, we divided health terms into 15 groups: those with 1 synset, those with 2 synsets and 1st health-related rank; other health terms with 2 synsets; ...; those with 7 synsets and 1st health -related rank; other health terms with 7 synsets; those with $\geq$ 8 synsets and 1st health-related rank; other health terms with $\geq$ 8 synsets.

PHI tweets with the PO terms	
Best Precision	0.774
Best F-score	0.652
PHI tweets without the PO terms	
Best Precision	0.738
Best F-score	0.649

Table 6: The best *F-score* and *Precision* of the PHI tweets extraction.

We computed *Precision* and *Fscore* of the extraction methods. Our empirical evidence showed that albeit the least ambiguous terms of synsets 1 and 2 give the highest *Precision*, the optimal *Fscore* is reached when the number of synsets reaches 6. Moreover, *Fscore*'s optimum at synset 6 is independent from the presence of personal ontology terms. In other words, it holds in both cases of personal health information extraction – with the PO terms and without them. Table 6 lists the best *F-score* and *Precision*. Note that our *Recall*= 100%.

The results showed that as the number of synsets associated with ontology terms increases, *Precision* of the extraction decreases but only slightly. This is an expected result of the word sense disambiguation, since, with more meanings associated with a given term, the more likely it is to be used in the non-health related contexts. This result, however, supported our premise of the importance of incorporating semantic information into the search.

4 Machine Learning of Tweets with PHI

On average, 200,000,000 tweets appear daily. [4] To be able to follow and extract tweets with PHI, we need to employ advanced automated software. In this section we show the advantage of using the

[3] http://wordnet.princeton.edu/

[4] https://blog.twitter.com/2011/200-million-tweets-day

Class	Relation to PHI	# of tweets
Class 1	the tweets with PHI	252
Class 2	tweets preceding PHI	251
Class 3	tweets following PHI	240

Table 7: Multi-class learning of PHI.

PHI ontology in Machine Learning of tweets containing PHI.

4.1 Classification problems

We apply classification technique to demonstrate that tweets with PHI are reliably differentiated from tweets without PHI if the extraction procedure used the PHI ontology. Hence, we classify the extracted tweets with PHI vs tweets without PHI. We use two types of tweets without PHI: a) those preceding the tweets with PHI, b) those following the tweets with PHI.

As a result, we state the learning experiments as a three-class classification problem. Classes are described in Table 7.

We applied Naive Bayes (NB) because of its reliable performance in previous Twitter classification studies (Bobicev et al., 2012).

4.2 Feature sets

Our next task was to define sets of words (i.e., features) that will represent tweets in classification. We contemplated between semantic PHI features and statistically selected features. We considered the use of semantic features to be undesirable. Semantic features were used to extract the tweets with PHI, thus representing tweets through them would bias an algorithm towards recognition of the tweets with PHI. On the other hand, we did not use the word statistic during the extraction procedure, thus, there would not be a pre-set classification bias if the features were selected statistically. Based on this consideration, we used four feature sets to represent the data:
Features I: all words with occurrence > 2;
Features II: words occur. > 2 that form the smallest subset of words which showed a better prediction of the class labels on the training set;
Features III: all words with occurrence > 5;
Features IV: words occur. > 5 that form the smallest subset of words which showed a better prediction of the class labels on the training set.

Three-class learning				
Features	*AUC*	*P*	*R*	*F*
I	**0.621**	0.459	0.448	0.452
II	0.569	0.386	0.388	0.386
III	0.607	**0.464**	**0.451**	**0.455**
I V	0.519	0.372	0.370	0.369
Baseline	0.497	0.115	0.339	0.172

Table 8: Classification of tweets with PHI. The best results are in **bold**.

4.3 Three-class learning

We used 10-fold cross-validation for the best classifier selection. We evaluated the performance by *Precision*, *Recall*, and *F-score*. Due to a relative imbalance of the data, we used *AUC* instead of a more traditional *Accuracy*. Also, *AUC*, representing a single point of the Reception Operating Characteristic curve, focuses on classifier's ability to avoid false classification (Sokolova and Lapalme, 2009).

Table 8 reports the average learning results. We computed baseline as the majority class classification.

The results show that classification beat the baseline on every feature set. The two-tailed t-test gives P equal to 0.067, 0.172, **0.064**, 0.220 on the four feature sets respectively. The most accurate identification of tweets with PHI happens when they are represented through words with occur. > 5, i.e., *P*, *R* and *F* are the highest. The most balanced identification of all the three classes happened on words with occur > 2, i.e. *AUC* is the highest. In the current case the feature selection substantially diminished the performance accuracy, unlike in previously reported studies of tweets with PHI(Bobicev et al., 2012).

We also wanted to know how well each class is differentiated among the three classes, depending on the features selected. Table 9 reports the classification results for each class separately.

We again see that the best identification of classes happens when the classifier can access words without any pre-selection. All the highest values but one were obtained on features representing words with occur. > 2 and > 5.

5 Related Work

We identify three major trends in mining for PHI on the Web.

Message boards In (Doing-Harris and Zeng-

Class I (Tweets with PHI)				
Features	*AUC*	*P*	*R*	*F*
I	**0.752**	0.607	**0.511**	0.555
II	0.624	0.426	0.458	0.442
III	0.700	**0.618**	0.508	**0.558**
I V	0.565	0.419	0.394	0.407
Class II (Tweets preceding PHI)				
Features	*AUC*	*P*	*R*	*F*
I	**0.580**	**0.408**	0.462	**0.433**
II	0.556	0.379	0.410	0.394
III	0.566	0.393	**0.470**	0.428
I V	0.500	0.347	0.414	0.377
Class III (Tweets following PHI)				
Features	*AUC*	*P*	*R*	*F*
I	0.531	0.362	**0.371**	0.366
II	**0.556**	0.351	0.295	0.32 0
III	0.554	**0.377**	**0.371**	**0.374**
I V	0.490	0.348	0.299	0.321

Table 9: Individual class recognition. The best results for each class are in **bold**.

Treiler, 2011), the authors extracted healthrelated terms from messages posted on PatientsLikeMe.com. To build a preliminary list of words, the authors applied entity recognition (dictionary look-ups, automated term recognition), N-gram modeling (frequency of consecutive words appearing in the messages) and symbolic processing (part-of-speech tagging and sentence parsing). User requests posted on an involuntary childlessness message board were studied (Himmel et al., 2009). In (Sokolova and Bobicev, 2011), the authors analyzed discussions about medications, treatment, illness and cure. Manual and automated methods were applied to recognize positive, negative and neutral opinions and positive and negative sentiments.

Blogsphere A keyword search was applied to the analysis of blogs written by military servicemen (Konovalov et al., 2010). The authors focused on finding terms that described clinically relevant combat exposure. In (Lagu et al., 2008), the authors manually examined blogs retrieved through Google searches medical blog, physician blog, doctor blog, nurse blog. The goal was to find blogs written by physicians or nurses that included some medical content (e.g., comments about health care system, laboratory studies).

Micro-blogosphere The occurrence of H1N1-related terms was studied in (Lampos and Christianini, 2010).The extraction method traced tweets that contained H1N1 and its synonyms (e.g., swine flu). Numerical evaluation of the methods' accuracy were reported by the authors of the both papers. Bobicev et al (2012) studied tweets that reveal PHI. However, their work was focused on sentiment analysis of these tweets.

6 Conclusions and Future Work

In this work, we have presented a mining method for personal health information in Twitter. We have shown that the use of the PHI ontology considerably improves PHI extraction if compared with other electronic resources of health information. We also have analyzed the impact of term meanings (WordNet) and general semantics (ontology of personal relations) on the extraction of PHI. We have demonstrated that semantic enhancement allows a reliable identification of messages with the topic of personal health.

We applied Machine Learning to demonstrate the advantage of our extraction method in classification of tweets with PHI. The need for classification arises because of a large amount of tweets appearing daily (approx. 200 mil. per day). A three-class classification had shown considerable improvement over the baseline results.

The presented work for mining Twitter messages is novel in several ways. First, it is specific to personal health information. Second, we incorporate health-related semantics into the mining process, and third, we build language patterns indicative for discussion of personal health information. To the best of our knowledge, there has not been a similar effort in mining information in Twitter.

Our future work includes text mining of lists of tweets posted by the same user (threads), analysis of the health information dissemination among the users. We will apply our approach on a considerably bigger set of the Twitter data. Finally, we aim to use posts from other social networks to look for similarities in the discussion of personal health on the Web.

Acknowledgments

This work was in part supported by NSERC Strategic Projects grant and CHEO Department of Surgery Research grant.

References

The title, the publication venue, the year.

Balicco, L., and Paganelli, C. 2011. Access to health information: going from professional to public practices Information Systems and Economic Intelligence: 4th International Conference - SIIE'2011

Bobicev, V., M. Sokolova, Y. Jafer, D. Schramm. Learning Sentiments from Tweets with Personal Health Information, Proceedings of Canadian AI 2012, Springer, 2012.

Chanlekha, H. and Collier, N. Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports, Journal of Biomedical Semantics, 1(3), 2010.

Chew, C. and G. Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS One, 5(11), 2010.

Doing-Harris, K. and Q. Zeng-Treiler. Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data. Journal of Medical Internet Research, 13(2):e37, 2011.

Ghazinour, K., S. Matwin, M. Sokolova. Monitoring and Recommending Privacy Settings in Social Networks, Proceedings of the 6th International Workshop on Privacy and Anonymity in the Information Society (PAIS 2013), p.p. 164 – 168, 2013.

Ghazinour, K., M. Sokolova, S. Matwin. Detecting Health-related Privacy Leaks in Social Networks Using Text Mining Tools, in Advances in Artificial Intelligence 26, Springer, 2013.

Himmel, W. and U. Reincke, H. Michelmann. Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums. Journal of Medical Internet Research, 11(3):e25, 2009.

Hersh W., Information retrieval: a health and biomedical perspective, 3rd ed., Springer, 2009.

Konovalov, S., M. Scotch, L. Post, C. Brandt. Biomedical Informatics Techniques for Processing and Analyzing Web Blogs of Military Service Members. Journal of Medical Internet Research, 12(4):e45, 2010.

Lagu, T., E. Kaufman, D. Asch, and K. Armstrong. 2008. Content of Weblogs Written by Health Professionals. Journal of General Internal Medicine, 23 (10): 1642–1646, 2008.

Lampos, V. and N. Christianini. "Tracking the flu pandemic by monitoring the social web". 2nd Workshop on Cognitive Information Processing, 2010.

Renahy, E. 2008. Recherche bd'infomation en matiere de sante sur INternet: determinants, practiques et impact sur la sante et le recours aux soins., Paris 6.

Sokolova, M. and G. Lapalme. "A Systematic Analysis of Performance Measures for Classification Tasks", Information Processing and Management, 45, p. 427–437, Elsevier, 2009.

Sokolova, M. and V. Bobicev. Sentiments and Opinions in Health-related Web messages. Recent Advances in Natural Language Processing, p.p. 132–139, 2011.

Sokolova, M. and D. Schramm. Building a patient-based ontology for mining user-written content. Recent Advances in Natural Language Processing, p.p. 758–763, 2011.

Sokolova, M., Jafer, Y., Schramm, D. "Text Mining for Personal Health Information on Twitter", Proceedings of IEEE HISB 2012

Sutton, C. and McCallum, A."An Introduction to Conditional Random Fields". Foundations and Trends in Machine Learning 4 (4), 2012.

Witten, I, E., Frank, M. Hall. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.