Analysing Relevant Diseases from Iberian Tweets

Víctor M. Prieto[1], Sergio Matos[2], Manuel Álvarez[1],
Fidel Cacheda[1], and José Luís Oliveira[2]

[1] University of a Coruña, Department of Information and Communication Technologies, Campus Elviña s/n, A Coruña, Spain
{victor.prieto,manuel.alvarez,fidel.cacheda}@udc.es
[2] University of Aveiro, DETI/IEETA, Campus Universitario de Santiago, 3810-193 Aveiro, Portugal
{aleixomatos,jlo}@ua.pt

Abstract. The Internet constitutes a huge source of information that can be exploited by individuals in many different ways. With the increasing use of social networks and blogs, the Internet is now used not only as an information source but also to disseminate personal health information. In this paper we exploit the wealth of user-generated data, available through the micro-blogging service Twitter, to estimate and track the incidence of health conditions in society, specifically in Portugal and Spain. We present results for the acquisition of relevant tweets for a set of four different conditions (flu, depression, pregnancy and eating disorders) and for the binary classification of these tweets as relevant or not for each case. The results obtained, ranging in AUC from 0.7 to 0.87, are very promising and indicate that such approach provides a feasible solution for measuring and tracking the evolution of many health related aspects within the society.

Keywords: Data mining, classification, social media, detecting health conditions.

1 Introduction

The Internet constitutes a huge source of information that can be exploited for various needs. For a long time, it has been used by individuals seeking medical information. However, with the advent of the Web2.0 paradigm, the Internet is now used not only as an information source but also to disseminate personal health information, experiences and knowledge [12] [15].

Much of this health related information is shared through social media platforms such as Twitter and Facebook. Twitter[1], for example, offers a micro-blogging platform that allows users to communicate through status updates limited to 140 characters, commonly referred to as "tweets". It has over 200 million active users[2], and around 400 million tweets are published daily. These large quantities of user generated content (UGC) can been exploited in different ways and represent great opportunities for data and text mining approaches in many fields of application. Mining these data provides

[1] http://www.twitter.com
[2] https://twitter.com/twitter/status/281051652235087872

M.S. Mohamad et al. (Eds.): *7th International Conference on PACBB*, AISC 222, pp. 69–76.
DOI: 10.1007/978-3-319-00578-2_10 © Springer International Publishing Switzerland 2013

an instantaneous snapshot of the public's opinions, and longitudinal tracking allows identification of changes in opinions [3]. This applies also to health related information, as can be verified by the various works that use Twitter and other user-generated data to assess and categorize the kind of information sought by individuals, to infer health status or measure the spread of a disease in a population. In [11], for example, the authors compared three web-based biosecurity intelligence systems and highlighted the value of social media, namely Twitter, in terms of the speed the information is passed and also because many issues or messages where not disseminated through other means. The greatest advantage of these methods over traditional ones is instant feedback: while health reports are published in a weekly or monthly basis, both tweets and query log of search engines can be obtained almost instantly. This characteristic is of extreme importance because early stage detection can reduce the impact of epidemic breakouts [1,7].

In this work, we propose an automated method, taking advantage of the wealth of data provided by Twitter, to measure the incidence of a set of health conditions in society, namely flu, depression, pregnancy and eating disorders. We focused our work on two official languages in the Iberian peninsula (Portuguese and Spanish), but the method we propose could be applied directly (e.g. South America) or adapted for other regions.

2 Related Work

Several works regarding the retrieval of health information from social media have already been published, with a major focus on measuring the incidence rate of influenza.

Chew and Eysenbach [3] suggested a complementary infoveillance approach during the 2009 H1N1 pandemic, using Twitter. They applied content and sentiment analysis to 2 million tweets containing the keywords "swine flu", "swineflu", or "H1N1". For this, they created a range of queries related to different content categories, and showed that the results of these queries correlated well with the results of manual coding, suggesting that near real-time content and sentiment analysis could be achieved, allowing monitoring large amounts of textual data over time. Signorini et al. [17] used Twitter to monitor public concern and levels of disease during the H1N1 pandemic in the United States. They collected tweets matching a set of 15 pre-specified search terms including "flu", "vaccine", "tamiflu", and "h1n1". They used content analysis to measure public interest and concern about this issue, and also applied support-vector regression to estimate influenza-like illness levels, using the Centers for Disease Control (CDC) data as reference. Using a model trained on 1 million influenza-related tweets, they reported average errors ranging from 0.04% to 0.93%. Lampos and Cristianini [10] and Culotta [5,6] also used regression models to estimate flu incidence rates in the United Kingdom and the United States respectively, obtaining correlation ratio of approximately 0.95. Aramaki et al. [1] applied SVM machine learning techniques to Twitter messages to predict influenza rates in Japan, achieving a correlation ratio of 0.89. Santos and Matos [14] combined data from Twitter and search engine logs in a

regression model to estimate the incidence of flu in Portugal, achieving a correlation ratio of 0.89.

Chunara et al. [4] analysed cholera-related tweets published during the first 100 days of the 2010 Haitian cholera outbreak. For this, all tweets published in this period and containing the word "cholera" or the hashtag "#cholera" were considered, and these data were compared to data from two sources: HealthMap, an automated surveillance platform, and the Haitian Ministry of Public Health (MSPP). They showed good correlation between Twitter and HealthMap data, and showed a good correlation (0.83) between Twitter and MSPP data in the initial period of the outbreak, although this value decreased to 0.25 when the complete 100 days period was considered.

Apart from analysing the incidence of flu and infectious diseases related events, the analysis of other health parameters using Twitter data has also been reported. Scanfeld et al. [15], for example, applied content analysis to 1000 tweets to explore evidence of misunderstanding or misuse of antibiotics. Heaivilin et al. [9] also applied content analysis to a set of 1000 tweets matching search criteria relating to dental pain. The content was coded using pre-established categories, including the experience of dental pain, actions taken or contemplated in response to a toothache, impact on daily life, and advice sought from the Twitter community. Bosleya et al. [2] analysed and categorized 60 thousand tweets concerning cardiac arrest and resuscitation, obtained during a 38 day period using a set of 7 search terms. All these works have in common that the content analysis is performed manually. This fact limits their application over long time periods, as well as great amount of data or large regions.

3 Method

In this article we propose a method to extract a set of tweets that show the presence of certain health conditions in people, as a point of the departure to infer the incidence of such conditions in society. This section describes the procedures used to obtain the tweets related to these conditions.

In order to obtain the different sets of tweets, we defined several regular expressions to extract only the tweets related to the studied diseases.

To create these expressions, we initially obtained a set of tweets containing the name of each condition, removing re-tweets and tweets that included links, and calculated the log-likelihood of the words that occurred within those datasets, therefore obtaining an ordered list of words associated with each disease. Based on these lists, on manual content analysis and on general knowledge about the studied diseases, we then defined the regular expressions for each specific condition.

Table 1 shows the regular expressions used to detect tweets related to the analysed diseases in Spanish. A similar list was used for Portuguese.

The use of regular expressions allowed obtaining large sets of tweets related to the specified diseases. However, among the obtained tweets, negative sentences that do not indicate the presence or absence of a disease in one person, such as "Hoping the flu does strike me again this winter", may also occur . To solve this problem, we applied machine learning techniques on the datasets obtained using the regular expressions, in order to filter such cases (see Section 3.1). This allowed differentiating the tweets that

Table 1. Regular expressions for detecting health disorders in Spanish tweets

Flu	Regular Expression
Flu	$(grip\|a-z\|+)$
Cold	$(costip[a-z]+)$
Flu Symptoms	$(fiebre.*grado(s)?)\|(grado(s)?.*fiebre)$
	$(((dolor(es)?\|(meduele)).*(cuerpo\|cabeza\|garganta)).*fiebre)\|$
	$(fiebre.*((dolor(es)?\|(meduele)).*(cuerpo\|cabeza\|garganta)))$
Pregnancy	
Pregnancy	$(embaraz[a-z]+)$
Common phrases	$(espero\|tendre).*((un\|una\|unos\|unas)?s(hij[a-z]+\|niño(s)?\|bebe(s)?\|niñit[a-z]+))$
	$(ser(e)?\|soy\|somos).*(padre(s)?\|madre)$
Depression	
Depression	$(depres[a-z]+)$
Depressed	$(deprim[a-z]+)$
Common phrases	$((problema(s)?\|disturbio(s)?).*(mental\|mentales\|psicologico(s)?\|psiquiatrico(s)?))$
	$(quiero).*(morir\|morir[a-z]+)$
	$(todo(s)?).*(dia(s)?).*(trist[a-z]+\|problema(s)?)$
Eating Disorders	
Obesity	$(obesidad\|obeso\|obesa)$
Overweight	$sobrepeso$
Bulimia	$(bulimia\|bulimica\|bulimico)$
Anorexia	$(anorexia\|anorexica\|anorexico)$
Bigorexia	$(vigorexia\|vigorexica\|vigorexica)$
Ebigorexia	$(ebigorexia\|ebigorexica\|ebigorexico)$
Orthorexia	$(ortorexia\|ortorexica\|ortorexico)$
	$(hacer\|hago\|hice).*((dieta(s)?\|regime(s)?)\|(dieta(s)?.*regime(s)?))$
	$(soy\|estoy\|mesiento).*gordo$
Common phrases	$((no)\|(quier[a-z]+)).*engordar$
	$(excesso\|problema\|riesgo\|peligro).*peso$
	$(enfermedad(es)?\|problema(s)?).*aliment[a-z]+$

Table 2. Number of features for each dataset

	Flu	Depression	Pregnancy	Eating Disorders
Spanish Tweets	608	721	698	567
Portuguese Tweets	842	983	1042	747

only mention a given disease from those which actually indicate that the person has the disease.

3.1 Machine Learning

To apply machine learning we need to obtain a set of features from the subsets of tweets related to the studied diseases (obtained by applying regular expressions (see Table 3). For that, we represented these tweets in a bag-of-words (BOW) model after removal of stopwords[3] and word stemming [13]. Character bigrams were also included in the feature set. According to the language and the disease studied, we obtained different sets of features, as shown on Table 2.

In order to identify the best classifier to our method, we have tested the obtained features with various machine learning techniques (SVM, Naïve Bayes, Decision Trees and Nearest Neighbour). To test these techniques, we used WEKA [8], an open source tool for data mining and machine learning that includes multiple implementations of different existing techniques.

[3] `http://snowball.tartarus.org/`

Table 3. Datasets used in the experiments

	Flu	Depression	Pregnancy	Eating Disorders
Spanish Tweets	827	3253	1985	412
Portuguese Tweets	1150	2845	2626	455

4 Experimental Results

In this section, we explain the datasets used in the experiments and several issues about how the experiments were made. We then analyse and discuss the results obtained with the proposed method for each disease.

4.1 Experimental Setup

To acquire the tweets for this study, we developed an application that uses the Twitter search API [18] and the geocoding information contained in the tweet metadata to obtain only tweets originated in Spain and Portugal. Furthermore, in order to filter out tweets not written in Spanish or Portuguese, we used the "language detector" library [16]. This library is based on Bayesian filters and has a precision of 0.99 in detecting the 53 languages it supports. Tweets were acquired during 30 days (from October 30th to November 30th, 2012).

The Spanish and Portuguese datasets contain approximately 5.8 and 4.5 million tweets, respectively. Table 3 shows the number of tweets considered for each language and disease pair, after applying the regular expressions shown in Table 1. The filtered tweets were manually labelled to be used for testing the machine learning algorithms. A tweet is considered true when it indicates the presence of one of the studied diseases in the person who have wrote the tweet. In any other case the tweet is considered false.

For the evaluation of the classifier we used a ten-fold cross validation technique. We used a polynomial kernel with $C = 1.0$, for SVM, and the default WEKA parameters for the remaining methods.

4.2 Results

Using all the features of each disease calculated for each country, we tested different implementations of the classifiers. The results for each type of classifier are shown on Table 4.

In the results shown, the Naïve Bayes classifier achieved the best results in all the cases except for 'Flu' in Spanish tweets. This classifier obtained in many cases a precision and a recall higher than 0.9, with an AUC always higher than 0.7, and often near 0.9. The second best classifier was the Decision Tree, followed by kNN. The worst results were obtained with the SVM classifier, with an AUC below 0.7 in some cases.

On the other hand, we can see that the best results were obtained in depression and in pregnancy (in Portuguese tweets). Regarding the country, in general, better results were obtained in the Portuguese dataset.

Table 4. Results obtained on the datasets. AUC = Area Under the receiver operating characteristic Curve.

		Spanish Tweets				Portuguese Tweets			
Disease	Classifier	F-Measure	Precision	Recall	AUC	F-Measure	Precision	Recall	AUC
Depression	Naïve Bayes	0.913	0.949	0.891	**0.878**	0.912	0.947	0.887	**0.833**
	SVM	0.946	0.948	0.944	0.739	0.902	0.934	0.876	0.691
	Decision Tree	0.976	0.968	0.985	0.845	0.974	0.963	0.985	0.762
	kNN	0.862	0.937	0.814	0.784	0.900	0.937	0.871	0.768
Pregnancy	Naïve Bayes	0.952	0.948	0.957	**0.703**	0.977	0.973	0.982	**0.877**
	SVM	0.940	0.942	0.939	0.644	0.945	0.975	0.920	0.679
	Decision Tree	0.947	0.944	0.951	0.689	0.978	0.971	0.985	0.801
	kNN	0.949	0.945	0.953	0.701	0.979	0.975	0.985	0.714
Flu	Naïve Bayes	0.766	0.759	0.775	0.743	0.667	0.667	0.669	**0.746**
	SVM	0.755	0.749	0.764	0.696	0.681	0.691	0.690	0.671
	Decision Tree	0.749	0.757	0.804	0.670	0.672	0.672	0.674	**0.746**
	kNN	0.761	0.756	0.799	**0.786**	0.687	0.687	0.689	0.745
Eating Disorders	Naïve Bayes	0.720	0.720	0.720	**0.714**	0.786	0.785	0.817	**0.744**
	SVM	0.683	0.688	0.679	0.607	0.725	0.729	0.720	0.650
	Decision Tree	0.785	0.756	0.817	0.630	0.869	0.838	0.902	0.690
	kNN	0.684	0.714	0.669	0.696	0.667	0.737	0.630	0.686

5 Conclusions

This article presents a method to extract a set of tweets that show the presence of certain diseases (flu, depression, pregnancy, eating disorder) in the society. The study was centred in Spain and Portugal, based on the geocoded data and on the language of the tweets. Using these sets of tweets we aim to measure the presence and evolution of a certain disease in society.

The proposed method is divided into two stages. First, we continuously gathered all tweets of each country and then filtered these tweets by means of several regular expressions, defined specifically for each disease. Secondly, we used machine learning methods, specifically Naïve Bayes, SVM, Decision Trees and kNN classifiers, in order to remove false positive documents identified with the regular expressions.

Compared to previous works, the main advantages proposed in this study are the detection of several health conditions in two distinct languages. The results obtained are very promising and indicate that such an approach provides a feasible solution for measuring and tracking the evolution of many health related aspects within the society.

Finally, we want to highlight the results obtained by our method applying Naive Bayes, which has obtained a precision and a recall close to 0.9. Based on this fact, we present Naive Bayes as the most suitable classifier for the proposed method to detect diseases in Twitter.

6 Future Work

Other types of user-generated content, such as Internet searches or comments to news articles, may also contain information related to some of these aspects. Thus, this information could be used to complement the data extracted from Twitter.

The proposed method may be extended to other languages and subjects, providing a continuous monitoring system of health pandemic or social issues, in a larger geographic region.

Acknowledgements. This research was supported by Xunta de Galicia CN2012/211, the Ministry of Education and Science of Spain and FEDER funds of the European Union (Project TIN2009-14203) and by "Fundação para a Ciência e a Tecnologia" (FCT, Portugal) under project PTDC/EIA-CCO/100541/2008 and Ciência2007 programme.

References

1. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568–1576. Association for Computational Linguistics (2011)
2. Bosley, J.C., Zhao, N.W., Hill, S., Shofer, F.S., Asch, D.A., Becker, L.B., Merchant, R.M.: Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication (2012)
3. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one 5(11), e14118 (2010)
4. Chunara, R., Andrews, J.R., Brownstein, J.S.: Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. American Journal of Tropical Medicine and Hygiene 86(1), 39–45 (2012)
5. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics, pp. 115–122. ACM (2010)
6. Culotta, A.: Detecting influenza outbreaks by analyzing Twitter messages, arXiv:1007.4748 [cs.IR] (2010)
7. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2009)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11, 10–18 (2009)
9. Heaivilin, N., Gerbert, B., Page, J.E., Gibbs, J.L.: Public health surveillance of dental pain via Twitter. Journal of Dental Research 90(9), 1047–1051 (2011)
10. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: 2010 2nd International Workshop on Cognitive Information Processing (CIP), pp. 411–416 (2010)
11. Lyon, A., Nunn, M., Grossel, G., Burgman, M.: Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. Transboundary and Emerging Diseases 59(3), 223–232 (2012)
12. Paul, M., Dredze, M.: You are what you tweet: Analyzing Twitter for public health. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 265–272 (2011)

13. Porter, M.F.: Snowball: A language for stemming algorithms. (published online, October 2001)
14. Santos, J.C., Matos, S.: Predicting Flu Incidence from Portuguese Tweets. In: Proceedings of IWBBIO 2013, Granada, Spain (March 2013)
15. Scanfeld, D., Scanfeld, V., Larson, E.L.: Dissemination of health information through social networks: twitter and antibiotics. American Journal of Infection Control 38(3), 182–188 (2010)
16. Shuyo, N.: Language detection library for java (2012)
17. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PloS One 6(5), e19467 (2011)
18. Twitter search api (2012), https://dev.twitter.com/docs/api/1/get/search (online; accessed November 20, 2012)