2017

Review of trends in health social media analysis

Liliya Akhtyamova

Mikhail Alexandrov

John Cardiff

Review of Trends in Health Social Media Analysis

Liliya Akhtyamova[1,2], Mikhail Alexandrov[2,3], and John Cardiff[1]
[1] Institute of Technology Tallaght, Dublin, Ireland
liliya.akhtyamova@postgrad.ittdublin.ie
john.cardiff@it-tallaght.ie
[2] Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia
[3] Autonomous University of Barcelona, Barcelona, Spain
malexandrov@mail.ru

***Abstract*—This paper surveys recent publications (2008-2017) on using social media data to study public health. The survey describes the main topics being discussed in forums and presents short information about methods and tools used for analysis health social media. We put especial attention on adverse drug reaction detection problem (ADR).**

Keywords—Health social media; review; adverse drug reactions; deep learning

INTRODUCTION

Social media is a modern phenomenon that has opened absolutely new possibilities for analysis of various aspects of a life of the human society in total or some group of people in particular. The medical domain is presented in various forums, where users discuss both general topics as the state of healthcare system or the specific questions concerning medicine, treatment etc. Such an information could be interesting to various governmental and private institutions. The former has an opportunity to evaluate the reaction of community on the laws and acts concerning healthcare as well as monitor the health condition of citizens and the latter can see a market for medicines for their production.

From the other hand, social media has provoked the development of NLP (Natural Language Processing), namely new models, methods and program systems. Medical domain presented in social media uses traditional approaches of NLP related to opinion mining. But not only these approaches. It considers specific problems related to diseases, treatment, social support of patients, and so on.

At the moment, there is a certain experience in analysis of data from health social media and in the development of NLP systems used in this analysis. The paper is organized as follows: section 2 gives retrospective of topics having been considered in recent publications related to health social media; section 3 presents methods for processing textual and numerical information; section 4 concludes the paper putting attention on future research.

TOPICS UNDER CONSIDERATION

Users of Health Social Media

There is an ongoing increase in the use of social media globally, including in health care contexts [1], [2].

Firstly, it is needed to define what is social media. According to the definition given by Kaplan and Haenlein [3] the social media is "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". They identified two components of social media: media-related (social presence, media richness) and social processes (self-presentation, self-disclosure). For the first one the most important things are intimacy and immediacy of the medium; how fast it facilitates human interaction (richness of the medium and the degree of social presence it allows); medium possess in the social media with the goal to resolve ambiguity and uncertainty. With respect to the social processes on social media, the people are possessed as having desire to express them, controlling for the impressions other people form of them; they wish to create an image of themselves. Usually, it is expressed through self-disclosure (e.g., thoughts, feelings, likes, dislikes).

With respect to personality traits, it was stated that woman and men are equally likely to engage, but only the men with less emotional stability were more frequent users of social media [4]. Gender is also an important factor with young cohort dominated over more mature users, while the last one is most important in terms of openness to new experience, especially in health domain. But the engagement in networking among seniors irresistibly increase. Today, more than 35% 65 and older report usage of social media, compared with just 2% in 2005 [5].

In this context, social media becomes important part of the heathcare, providing speed and rather different way of interaction between individuals and health organizations. The healthcare professionals, as well as general public, patients use social media to retrieve medical information

and communicate about health issues [1]. According to MedTechMedia[1], 31% healthcare professionals are engaged in professional networking on social media. According to statistics, 18-24 years olds are more likely to twice as often seek for health-related information on social media, than 45 to 54 years olds[2], while 90% of the first ones trust medical information shared by others on social media network[3]. Moreover, 41% of people said social media will affect their choice of treatment, clinic and doctor[4].

All these makes social media irresistibly important source for health-related research.

ADR in Health Social Media

Adverse drug reactions (ADRs) are proved to be the reason of serious injuries and dies of more than 700000 people in USA [6]. One of the reasons is the limited possibilities of the preliminary clinical observations to predict all consequences. So, it is necessary to have any additional ways to detect side-effects of the medications, such as antidepressants, monoamine oxidase inhibitors, etc. One should say that ADRs has essential economic effect [7], namely, about 5.3% of hospital admissions are associated with ADRs [8].

Therefore, revealing adverse drug reactions is of paramount importance for the government authorities, drug manufacturers and patients. Data from the European Medicines Agency (EMA) shows that patients are not reporting side effects adequately through official channels, making it necessary to explore some different ways of ADR monitoring.

To attract researchers who explore automatic methods for the analysis of social media data to the problem of public health surveillance and monitoring, different workshops and sessions are organized worldwide. So, in the paper [9], the results of recent session “Social Media Mining for Public Health Monitoring and Surveillance" are published. Many interesting cutting-edge researches in the area of public health were highlighted. For example, an Instagram -- an increasingly popular social media platform where people share photos and videos -- was proven to be useful source of information for ADRs and drug interaction detection [10]. In other research [11] the authors try to categorize products (nutritional supplements) based on the users reviews on the popular online purchase platform Amazon.com to reveal possible danger of these products.

Another popular platform, which is worth to mention is Twitter. As of 1st September 2016, Twitter has over 342,000,000 active users and grows by 135,000 users every day, generating daily 58,000,000 tweets[5]. In this context, Twitter platform is advantageous over others as provides open API for research purposes to download their data, which could be very helpful for the task of ADR revealing and monitoring.

Other topics

Other directions of text analysis for medicine include medical concept extraction task for detecting mention of proteins, genes, drugs, diseases, tests and treatments and relation extraction tasks for extracting relations from texts about drug-drug interactions, drug-symptoms-treatment relations and etc. [12], [13].

Especially important source of medical information became forums. Examples include the prediction of whether people will stay on or leave health forums (such as DailyStrength[6] and HealthBoards[7]) and investigating why they do so [14]. This has shown to be a promising area, as continued participation in such kind of forums can be very fruitful for both patients and doctors. Other examples include pandemic tracking (for example, utilizing Twitter to track the discussion splash in different regions of USA during influenza A (H1N1) pandemic to prevent panic among citizens [15]; utilizing smoking cessation patterns on Facebook [16], as well as organizing different anti-smoking and other campaigns, which goal is to raise awareness among different cohorts of people (especially the youngsters) on essential world health-related problems[8]; revealing drug abuse [17] and monitoring malpractice [18] on Twitter.

Moreover, social media can provide researchers with specific kinds of information that is usually unavailable due to data protection legislation, including a person’s age, nationality, sex, and geolocation. It also helps to reveal users’ habits and interests, all of which can play a part in diagnostics and early detection of the different health disorders.

METHODS

Methods for Text Mining

In the context of Health Social Media analysis all the standard Machine Learning tools can be applied, however, some of them can perform poorer due to the specific features of health-related texts. Overall, there are two different tasks in processing texts: classification and different extraction tasks. Any existing supervised learning methods can be applied to classification, e.g., naive Bayesian classification, support vector machines (SVM), Maximum Entropy. The majority of approaches applied to the extraction tasks are feature or kernel based, identifying terms using a list of precompiled words and different rules and syntactic information about the sentences [19]–[21]. However, the performance of all described methods is highly dependent on suitable feature set selection, which is not only tedious and time-consuming task but also require domain knowledge and is dependent on other NLP systems. The solution for this problem is approaches, based on discovering patterns in texts. The most powerful way of doing that is neural networks which are becoming increasingly popular, and they have shown promising performances in medical image analysis [22], as well as in many NLP-related

[1] www.medtechmedia.com
[2] http://www.adweek.com
[3] https://searchenginewatch.com
[4] http://thesparkreport.com
[5] http://www.statisticbrain.com/twitter-statistics

[6] https://www.dailystrength.org/
[7] http://www.healthboards.com/
[8] see, for example, http://www.bangthetable.com/public-health-social-media-campaigns/

problems [23]–[25]. The most common tools for deep learning tasks are TensorFlow and Theano libraries, which could be fine-tuned to the user needs. Most of the deep neural networks in NLP utilise an embedding that projects each unique word into a dense lower-dimensional space (typically from 50 to 300 dimensions) and use it as the input of the network. The most popular tool for word embeddings construction is word2vec, which could be pretrained on a huge corpus of unlabeled data. For example, in a paper [26] the authors focused on extracting relations from clinical discharge summaries using domain invariant convolutional neural network. Their results outperformed the best result of baseline model on the i2b2-2010 clinical relation extraction challenge dataset.

ADR Detection

There is a number of papers, which contribute to the problem of ADR detection. The earliest work on ADR extraction was [27], where the authors investigated the potential of user comments for early detection of unknown ADR. This and the subsequent works were mainly focused on a limited number of drugs and target ADR and were based on hand-designed features [17], [20], [28].

Nevertheless, some improvements were made. For example, in a recent paper [29] the authors made classifier based on users estimations from the AskaPatient forum. They used an ensemble of many CNNs, which showed to be effective on this kind of data. Another work contributing to the topic of ADR detection is [30], where authors used a publicly available ADE corpus for binary medical case reports classification. In [17] the authors tried to combine ADE corpus with data from social media (DailyStrength, Twitter). They concluded that significant improvements in performance can be achieved by coupling Twitter and DailyStrength datasets, while combining DailyStrength and Twitter datasets with ADE training data didn't give notably better results. In another paper on the same dataset [31] the authors built different sophisticated CNNs models, with the best result achieved by simple CNN, fed with the additional set of features.

A recent challenge organized by the DiegoLab research lab aimed to tackle the deficit of research on the topic of ADR detection in medicine. The authors constructed a dataset using Twitter data, and labeled more than 10,000 tweets which contained the names of the 74 top selling drugs in 2013, including their misspellings. Approximately 20% of these mentioned ADRs. The unbalanced nature of the dataset became the main problem for the teams participating in this competition.

The best result was achieved by [32]. They conducted their experiment, using Support Vector Machines with LibSVM implementation[9]; The results they got were ADR F-score 0.597, non-ADR F-score 0.943 and accuracy 90.1%. The main disadvantage of their work became the huge number of features (from sentiment analysis, polarity classification, topic modeling) was needed to obtain to get this result with additional corpora from health forum and medical records added to train the model. Later the same or even better results were obtained with neural networks. For example, in a paper [33] the authors tried to use a simple neural network model. The results obtained were ADR F-score 0.43; non-ADR F-score 0.943; accuracy 90.1%. In another paper [31], different deep neural network algorithms were implemented for this task. Although more sophisticated algorithms were represented, neural networks demonstrated the best result with 51% of ADR F-score. Later in another paper [34] even better results (ADR F-score 68%), using semi-supervised CNNs.

For the relation extraction task, the good result shows [26] with their domain invariant convolutional neural network for the drug-drug interaction extraction task from biomedical texts. Their experimental results on the SemEval-2013 DDI extraction dataset showed that their Joint AB-LSTM model outperforms all the existing methods, including those relying on handcrafted features.

[9] http://www.csie.ntu.edu.tw/?cjlin/libsvm/

CONCLUSIONS

This paper surveys the field of health social media mining. Due to the increasing awareness of people, government and different authorities about health issues and a wide variety of practical applications, it becomes a very active research area. We presented a general topic of health social media and its participants, followed by the problem of ADR detection and an overview of other important implications of social media mining for medical purposes. We then described some methods, paying special attention to the novel deep learning methods, which recently demonstrated their advantages over the other existing models using handcrafted features. The future improvements need more linguistic resources to be included to the methods related to processing texts from health social media.

REFERENCES

[1] R. Thackeray, B. L. Neiger, C. L. Hanson, and J. F. McKenzie, “Enhancing Promotional Strategies Within Social Marketing Programs: Use of Web 2.0 Social Media,” Health Promot. Pract., vol. 9, no. 4, pp. 338–343, Oct. 2008.

[2] C. L. Ventola, “Social media and health care professionals: benefits, risks, and best practices.,” P T, vol. 39, no. 7, pp. 491–520, Jul. 2014.

[3] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” 2007.

[4] T. Correa, A. W. Hinsley, and H. G. de Zúñiga, “Who interacts on the Web?: The intersection of users’ personality and social media use,” Comput. Human Behav., vol. 26, no. 2, pp. 247–253, Mar. 2010.

[5] A. Perrin, “65% of adults now use social networking sites – a nearly tenfold jump in the past decade,” Pew Res. Cent., 2015.

[6] R. A. Lehne and L. D. Rosenthal, Pharmacology for Nursing Care. Elsevier Health Sciences, 2013.

[7] J. Sultana, “Clinical and economic burden of adverse drug reactions,” J. Pharmacol. Pharmacother., vol. 4, 2013.

[8] C. Kongkaew, P. R. Noyce, and D. M. Ashcroft, “Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies,” 2008.

[9] M. J. Paul et al., “Social Media Mining for Public Health Surviellance and Monitoring,” Pacific Symp. Biocomput., 2016.

[10] R. B. Correia, L. Li, and L. M. Rocha, “Monitoring Potential Drug Interactions and Reactions via Network Analysis of Instagram User Timelines,” Oct. 2015.

[11] R. Sullivan, A. Sarker, K. O ’Connor, A. Goodin, M. Karlsrud, and G. Gonzalez, “Finding Potentially Unsafe Nutritional Supplements from User Reviews with Topic Modeling,” Pacific Symp. Biocomput., 2016.

[12] M. Craven, “Constructing Biological Knowledge Bases by Extracting Information from Text Sources,” ISMB-99 Proc., 1999.

[13] R. Mcdonald and F. Pereira, “Identifying gene and protein mentions in text using conditional random fields,” BMC Bioinformatics, vol. 6(Suppl 1), 2005.

[14] F. Sadeque, T. Pedersen, T. Solorio, P. Shrestha, N. Rey-Villamizar, and S. Bethard, “Why Do They Leave: Modeling Participation in Online Depression Forums,” pp. 14–19, 2016.

[15] C. McNab, “What social media offers to health professionals and citizens.,” Bull. World Health Organ., vol. 87, no. 8, p. 566, Aug. 2009.

[16] L. L. Struik and N. B. Baskerville, “The Role of Facebook in Crush the Crave, a Mobile- and Social Media-Based Smoking Cessation Intervention: Qualitative Framework Analysis of Posts,” J. Med. Internet Res., vol. 16, no. 7, p. e170, Jul. 2014.

[17] A. Sarker et al., “Utilizing Social Media Data for Pharmacovigilance: A Review HHS Public Access,” J Biomed Inf., vol. 54, pp. 202–212, 2015.

[18] A. Nakhasi, R. J. Passarella, S. G. Bell, M. J. Paul, M. Dredze, and P. J. Pronovost, “Malpractice and Malcontent: Analyzing Medical Complaints in Twitter,” 2012.

[19] A. Benton et al., “Identifying potential adverse effects using the web: A new approach to medical hypothesis generation,” J. Biomed. Inform., vol. 44, no. 6, pp. 989–996, Dec. 2011.

[20] M. Yang, X. Wang, and M. Kiang, “Identification of Consumer Adverse Drug Reaction Messages on Social Media,” PACIS 2013 Proc., 2013.

[21] S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep, and R. Srinivasan, “A pipeline to extract drug-adverse event pairs from multiple data sources,” BMC Med. Inform. Decis. Mak., vol. 14, no. 1, p. 13, Dec. 2014.

[22] U. K. Sikdar and B. Gambäck, “Feature-Rich Twitter Named Entity Recognition and Classification,” pp. 164–170, 2016.

[23] A. Severyn and A. Moschitti, “Twitter Sentiment Analysis with Deep Convolutional Neural Networks,” 2015.

[24] N. Santos and M. Gatti, “Deep convolutional neural networks for sentiment analysis of short texts,” COLING, pp. 69–78, 2014.

[25] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv Prepr. arXiv1408.5882, 2014.

[26] S. Kumar Sahu, A. Anand, K. Oruganty, and M. Gattu, “Relation extraction from clinical texts using domain invariant convolutional neural network,” Proc. 15th Work. Biomed. Nat. Lang. Process., no. October, 2016.

[27] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez, “Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks,” Proc. 2010 Work. Biomed. Nat. Lang. Process. Assoc. Comput. Linguist., pp. 117–125, 2010.

[28] J. Bian, U. Topaloglu, and F. Yu, “Towards large-scale twitter mining for drug-related adverse events,” in Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB ’12, 2012, p. 25.

[29] L. Akhtyamova, A. Ignatov, and J. Cardiff, “A Large-Scale CNN Ensemble for Medication Safety Analysis,” Nat. Lang. Process. Inf. Syst. NLDB 2017. Lect. Notes Comput. Sci., vol. 10260, 2017.

[30] H. Gurulingappa, A. Mateen-Rajput, and L. Toldo, “Extraction of potential adverse drug events from medical case reports.,” J. Biomed. Semantics, vol. 3, no. 1, p. 15, Dec. 2012.

[31] T. Huynh, Y. He, A. Willis, and S. Rüger, “Adverse Drug Reaction Classification With Deep Neural Networks,” Proc. COLING 2016 Tech. Pap. COLING, pp. 877–887, 2016.

[32] A. Sarker and G. Gonzalez, “Portable automatic text classification for adverse drug reaction detection via multi-corpus training,” J. Biomed. Inform., vol. 53, pp. 196–207, 2015.

[33] B. Pastel and B. Villanueva, “Discovering Adverse Drug Reactions via Natural Language Processing of Twitter Posts,” 2015.

[34] K. Lee et al., “Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks,” in Proceedings of the 26th International Conference on World Wide Web - WWW ’17, 2017, pp. 705–714.