RUNNING HEAD: MACHINE LEARNING ON HIV DATA

Toward Automating HIV Identification: Machine Learning for Rapid Identification of HIV-related Social Media Data

Sean D. Young[1,2], Wenchao Yu[2,3], Wei Wang[2,3]

Correspondence to:

Sean Young, PhD, MS

University of California, Institute for Prediction Technology

10880 Wilshire Blvd., Suite 1800

Los Angeles, CA 90024

Phone: 310-794-0619 (Ext. 240)

Fax: 310-794-2768

Email: sdyoung@mednet.ucla.edu

E-mail addresses of authors:

SY: sdyoung@mednet.ucla.edu

WY: yuwenchao@ucla.edu

WW: weiwang@cs.ucla.edu

This work was funded by the UCLA CFAR Grant 5P30 AI028697, the UCLA AIDS Institute, and the National Institutes of Health (Grant No: U01:5U01HG008488; RO1: MH106415-01) and by the University of California Office of the President (CA-15-329077).

[1] Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, United States of America

[2] University of California Institute for Prediction Technology, University of California, Los Angeles, CA, United States of America

[3] Department of Computer Science, University of California, Los Angeles, CA, United States of America

ABSTRACT

Introduction: “Social big data” from technologies like social media, wearable devices, and online searches continue to grow and can be used as tools for HIV research. Although researchers can uncover patterns and insights associated with HIV trends and transmission, the review process is time-consuming and resource intensive. Machine learning methods derived from computer science might be used to assist HIV domain experts by learning how to rapidly and accurately identify patterns associated with HIV from a large set of social data.

Methods: Using an existing social media dataset that was associated with HIV and coded by an HIV domain expert, we tested whether four commonly used machine learning methods could learn the patterns associated with HIV risk behavior. We used the 10-fold cross validation method to examine the speed and accuracy of these models in applying that knowledge to detect HIV content in social media data.

Results and Discussion: Logistic regression and random forest resulted in the highest accuracy in detecting HIV-related social data (85.3%) while the Ridge Regression Classifier resulted in the lowest accuracy. Logistic regression yielded the fastest processing time (16.98 seconds).

Conclusion: Machine learning can enable social big data to become a new and important tool in HIV research, helping to create a new field of “digital HIV epidemiology.” If a domain expert can identify patterns in social data associated with HIV risk or HIV transmission, machine learning models could quickly and accurately learn those associations and identify potential HIV patterns in large social datasets.

Key Words: HIV prevention, social data, big data, machine learning, epidemiology

INTRODUCTION

HIV remains a significant public health problem, with more than 1 million people infected and approximately 14% of infected individuals in the United States unaware of their HIV status.[1] Early and rapid identification of risk behaviors could help to reduce the spread of HIV and prevent outbreaks. HIV researchers, including state and local health departments, commonly use surveys, interviews, and HIV diagnoses data as surveillance tools to monitor HIV epidemiology and risk behaviors.[2] Although these methods are established and have been generally successful, they also have limitations, including 1) a lag time in reporting, as the tools provide data on behaviors and diagnoses that have already occurred, and typically report these cases months or years after they occur; 2) data quality issues, as existing tools can have missing and/or biased data; and 3) extensive time involvement, as these methods require researchers to engage in a substantial number of hours collecting, analyzing, and reporting data.[1,2] Therefore, innovative approaches are needed to supplement existing tools and improve HIV surveillance systems.

Social media use has been rapidly growing,[3] and research has already shown that data from these technologies can be used for novel approaches to public health surveillance, including the identification of HIV-related risk behaviors.[2,4] Social media and social networking platforms, such as Facebook and Twitter, allow users to easily and freely communicate with each other by sharing pictures, short messages, website links, and other forms of multimedia communication. These sites provide an enormous amount of data (e.g., approximately 500 million tweets per day on Twitter).[5] Social media data are quickly becoming a core part of biomedical big data because they can be modeled alongside other biomedical data and used to predict health and medical outcomes.[6–8]

Recent work has already shown the feasibility of using data from the real-time social media site, Twitter, to identify and track trends where sexual risk behaviors might be occurring in order to detect potential HIV outbreaks.[2,9] For example, we showed that: 1) tweets can be extracted and identified to suggest that people plan to or currently engage in sex- and drug-related risk behaviors; 2) tweets suggesting the occurrence of these behaviors can be mapped to indicate their origin; and 3) these data can be merged and modeled alongside U.S. statistics of actual HIV cases.[2] However, the process for identifying tweets associated with HIV risk behaviors is time-intensive, susceptible to human error, and cannot be implemented by public health researchers on a large scale: it requires a domain expert in HIV epidemiology and/or psychology to manually view and code tens of thousands of tweets to determine whether they may be associated with sexual- or other HIV-related risk behaviors.

Modern computer science methods have allowed computers to "learn" the skills of human domain experts and apply these skills to a broad range of applications, such as monitoring trends in influenza transmission,[10] predicting crime rates,[11] and forecasting the stock market.[12] These machine learning approaches can be applied to HIV surveillance to analyze social media big data, or more broadly to learn patterns in data identified by an HIV domain expert and apply those patterns to large datasets. Specifically, the following steps could be implemented: 1) an HIV expert, such as an HIV epidemiologist or psychologist, could identify social data (e.g., tweets) associated with potential HIV risk, 2) the HIV expert could "teach" their process of identifying these data to a machine by supplying the machine with HIV-coded (labeled) data, and 3) the machine could learn this coding method and ultimately apply it to a large set of unlabeled social data. However, no studies have been conducted on this process, making it unclear whether

and how machine learning methods can be used to simplify social data analysis so that it can become a useful tool in HIV research.

The objective of this study was to identify whether machine learning approaches can be used to assist a domain expert who manually searches social data for HIV-related content. The machine learning tool would be able to extract free-text social data (e.g., tweets), learn from the coding of a domain expert who identifies HIV-related tweets, and automate this process on a large scale so that it could be conducted rapidly. In order for a machine learning tool to be incorporated into public health settings, the method would need to be both accurate and fast. We therefore sought to test a variety of machine learning methods to determine whether a machine could be given HIV-related tweets and learn how to quickly and accurately use a training dataset to identify new HIV-related social media content.

METHODS

We used a dataset from a previously published paper that contained 6,387 English-language tweets from the United States.[2] Tweets were collected between May 26, 2012, and December 9, 2012, using Twitter's free Advanced Programming Interface (API). In the study, a researcher identified whether the tweets were related to HIV risk behaviors, and ultimately associated with HIV cases. A tweet was classified as HIV-related if it contained a word or pattern of words (associations between words) related to sexual risk–related attitudes and behaviors, and/or HIV-related substance use (e.g., stimulants or opiates that have been shown to be associated with HIV). Tweets that did not contain information related to sexual, drug, or HIV risk were determined to "not be associated with HIV risk behaviors." Tweets were reviewed to determine that they were in fact related to HIV (e.g., by being related to sexual and/or drug-

related risk behaviors).[13] Overall, 2,191 (34%) tweets were identified as related to HIV risk behaviors. These tweets were associated with HIV cases at the county level and were therefore used as a training set of HIV-related tweets that would be learned by a computer.

To determine whether a machine could learn to identify HIV-related tweets, we developed and tested the accuracy of four widely used machine learning methods on the Twitter dataset: logistic regression, random forest, support vector machine (SVM), and ridge regression classifier. Logistic regression is a classification method used to estimate the probability of a binary label (HIV-related tweets vs. non-HIV–related tweets) based on features.[14] Random forest is an ensemble learning method that can be utilized for classification tasks.[15] It constructs a multitude of decision trees and makes a prediction based on the mode of the classes of the individual trees; 100 decision trees were used in our random forest model. SVM is a supervised learning model that learns a classifier to separate categories by a clear gap that is as wide as possible.[16] In our experiment, a linear SVM was applied. Ridge regression classifier imposes a penalty on the size of coefficients of ordinary least squares problem. The ridge coefficients minimize a penalized residual sum of squares.[17]

Data Analysis

The goal of our analysis was to determine whether, when given a training set of HIV-related tweets, machine learning models could accurately identify similar patterns in a new dataset of tweets. The tweets were filtered to extract relevant and important features prior to applying the machine learning models. According to standard machine learning methods, English stop words (e.g., “the,” “and,” “that”) were removed as these words do not significantly contribute to the accuracy of the machine learning models.[18] Term frequency-inverse document

frequency (TF-IDF) was applied to determine the importance of words in the tweets while adjusting for the fact that some words appeared more frequently. The bag-of-words model, a natural language processing method, was applied to create a word dictionary. This method learns a word from the tweets, then identifies the frequency of the word that appears in all tweets. The most frequently used words were determined using the P-Growth method.[19]

A 10-fold cross validation was applied to the dataset following the filtering process. Cross validation aims to find the model with the highest generalization ability.[20] Using 10-fold cross validation enabled us to divide the entire dataset (N = 6,387 tweets) into 10 randomly selected subsets of data. This method allowed us to train a model and then test it on a new set of data to ensure the patterns were correctly identified. The overall accuracy of the model for each machine learning method was calculated based on averages of the accuracies from 10 rounds of cross validation. In each machine learning method used, we also explored different types of features that could increase prediction accuracy, including word compounds, user features, content features, network features, and hashtag features.[21] The frequent pattern mining approach was also used to extend the dictionary with word compounds.

RESULTS

The bag-of-word model identified 9,111 words. The average accuracy of the 10-fold cross validation for the machine learning methods using word counts, word compounds, user features, content features, network features, and hashtag features is shown in Table 1.

[Insert Table 1]

Logistic regression and random forest plot resulted in the highest accuracy, while the ridge regression classifier resulted in the lowest accuracy. Logistic regression yielded the fastest processing time (Table 2).

[Insert Table 2]

DISCUSSION

To our knowledge, this is the first study to suggest that machine learning on social data can be used to assist HIV epidemiologists and public health researchers in addressing HIV-related outcomes. Current state-of-the art epidemiologic methods like surveys and case reporting are being successfully used by health departments and researchers, but this study presents data on how health departments can use social data as an additional tool to address HIV prevention, testing, and treatment efforts.

Although recent research has provided initial support to suggest that social data can be used to monitor and predict HIV outcomes, researchers are unclear which methods can be used to analyze these data due to the large and growing volumes of social data.[2,9] The current study provides a suggested method that can be used to integrate machine learning models with the expertise of an HIV researcher or epidemiologist. That is, if an HIV domain expert can identify content in social media that is associated with HIV risk behaviors, then machine learning methods appear to be able to rapidly and accurately learn from the domain expert and apply this knowledge to identify HIV risk behaviors in a large dataset. This process, therefore, has the potential to improve HIV research by allowing an HIV expert to work in tandem with a computer scientist who can help facilitate large-scale implementation of the domain expert's knowledge. In

our study, we found that the logistic regression method resulted in the highest overall performance.

Importantly, accuracy alone is not the only important attribute in determining whether public health researchers can incorporate machine learning tools to automate their work. Although all four of the models had fairly high accuracy (~85%), the random forest model took more than 40 times as long to process as the fastest model, logistic regression. Although a longer processing time of 700 seconds might not seem like much time, processing time can become days or weeks when instead of analyzing thousands of tweets a computer needs to analyze billions of social media posts. As researchers start studying “big data” related to HIV and other public health issues, one of the emerging issues is whether it will be possible to rapidly analyze large amounts of data. For example, a machine that can be used to identify HIV-related tweets in real-time will be much more useful to a public health organization if it can identify those tweets in real time rather than days or weeks after the occurrence of the tweet. As data continues to accumulate, such as with social media data, it will be increasingly important that tools are available that can quickly process large amounts of data. This study found that machine learning tools can analyze a large amount of data points and provide insights about HIV in a few minutes.

This study was limited by the use of an existing training dataset. The training data were found to be associated with HIV cases and therefore used as a dataset of HIV-related tweets; however, there is no way to validate whether the tweets were actually related to HIV. It is common in epidemiology, health economics, and other social and health sciences to discover an association that makes intuitive sense and then to further investigate that association, which was the approach taken in the present study. Broadly, the findings of this study are important because they signify that researchers can provide a machine with social media data with a valid

relationship to HIV ("gold standard data") and apply machine learning methods to rapidly learn those patterns and apply them to a large new dataset. Finally, although not a limitation, it is important to note that just because tweets were found to have been associated with HIV-related risk behavior does not mean that the person tweeting is necessarily going to act on that behavior or is at risk for HIV. By identifying tweets associated with HIV risk behaviors, it is therefore important to ensure that the tweets are not linked back to these individuals as that could lead to stigmatization. For that reason, this paper presents a partial list of commonly used keywords but does not provide data on the actual tweets in order to prevent identifying the individuals who made those tweets. A large and growing area of research will be focused on how to address the logistical and ethical issues associated with social data.[22–24]

Although current epidemiologic tools such as surveys and case reporting have limitations, machine learning on social data also has limitations. For example, social data are only available if people choose to use social technologies and allow others to access their data. This issue can lead to a biased participant sample. While quantitative or scale-based questionnaire items limit the depth of information a person can provide, character limits in some social technologies similarly limit the depth of information people can provide. The purpose of this paper is therefore not to convey that social data is a standalone tool, but rather to show that is can be used as an additional tool in the "surveillance toolbox" as well as to present a method for how to use these data. We hope that public health departments will begin to consider approaches from this new and growing field of "digital HIV epidemiology" and learn how social data can be used to help monitor HIV-related trends. We believe that digital HIV epidemiology analyses might be used to augment current tools and help healthcare personnel act quickly to deliver interventions, as needed.

CONCLUSION

This study is the first we know of to explore whether and how modern machine learning methods can be used to learn HIV associations found in social data and apply that knowledge to a new set of social data. As the body of social media data continues to grow, it will provide a rich source of information that can be used to assist public health researchers monitor, predict, and prevent the spread of HIV. Consequently, methods such as machine learning that can rapidly and accurately be used to improve the work of public health researchers will become increasingly important.

Competing interests

None to declare

Acknowledgements

We wish to acknowledge Sam Liu for his feedback on an early draft of the manuscript.

Author contributions

S.Y. contributed to the study design, data collection, data analysis and interpretation, and writing of all drafts of the manuscript. W.Y. contributed to data collection, data analysis, data interpretation, and writing of the manuscript. W.W. contributed to the study design, data interpretation, and writing of the manuscript.

REFERENCES

1. Monitoring Selected National HIV Prevention and Care Objectives by Using HIV Surveillance Data. Centers for Disease Control and Prevention website. Available at: http://www.cdc.gov/hiv/library/reports/hiv-surveillance/hiv-surveillance-supplement-17-3.html. Accessed October 11, 2016.

2. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med*. 2014;63:112-115.

3. Duggan M, Ellison NB, Lampe C, Am, Lenhart A, Madden M. Social media update 2014. Available at: http://www.pewinternet.org/2015/01/09/social-media-update-2014/. Accessed August 1, 2016.

4. Young SD. Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends Microbiol*. 2014;22(11):601-602.

5. Twitter Usage Statistics. Internet Live Stats website. Available at: http://www.internetlivestats.com/twitter-statistics/. Accessed October 11, 2016.

6. Signorini A, Segre AM, Polgreen PM. 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*. 2011;6(5):e19467.

7. Gore RJ, Diallo S, Padilla J. You are what you tweet: connecting the geographic variation in America's obesity rate to twitter content. *PloS One*. 2015;10(9):e0133505.

8. Choudhury MD, Choudhury MD, Counts S, Horvitz E, Gamon M. Predicting depression via social media. Microsoft website. Available at: https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/. Accessed October 11, 2016.

9. Ireland ME, Schwartz HA, Chen Q, Ungar LH, Albarracín D. Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychol*. 2015;34(Suppl):1252-1260.

10. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS One*. 2013;8(12):e83672.

11. Gerber MS. Predicting crime using Twitter and kernel density estimation. *Decis Support Syst*. 2014;61:115–125.

12. Bollen J, Mao H, Zeng X-J. Twitter mood predicts the stock market. *J Comput Sci*. 2011;2(1):1-8.

13. Shoptaw S, Reback CJ. Associations between methamphetamine use and HIV among men who have sex with men: a model for guiding public policy. *J Urban Health Bull N Y Acad Med*. 2006;83(6):1151-1157.

14. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol*. 1958;20(2):215-242.

15. Ho T. Random decision forests. In: *Proceedings of the Third International Conference IEEE*. IEEE;1995:278-282.

16. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 20(3):273-297.

17. Draper N, Smith H, Pownell E. *Applied regression analysis*. New York: Wiley.

18. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Available at: https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining. Accessed October 11, 2016.

19. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. New York, NY: ACM; 2000:1–12.

20. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2*. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995: 1137–1143.

21. Kong S, Mei Q, Feng L, Zhao Z, Ye F. On the real-time prediction problems of bursting hashtags in Twitter. Available at: https://www.researchgate.net/publication/259624858_On_the_Real-time_Prediction_Problems_of_Bursting_Hashtags_in_Twitter. Accessed October 11, 2016.

22. Grajales III FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: A review and tutorial of applications in medicine and health care. *J Med Internet Res*. 2014;16(2):e13.

23. Curtis BL. Social networking and online recruiting for HIV research: ethical challenges. *J Empir Res Hum Res Ethics*. 2014;9(1):58-70.

24. Chiu CJ, Menacho L, Fisher C, Young SD. Ethics issues in social media–based HIV prevention in low- and middle-income countries. *Camb Q Healthc Ethics*. 2015;24(3):303-310.

TABLE 1. Accuracy of the machine models using word counts, word compounds, user features, content features, network features, and hashtag features as predictive features (N = 6,387 tweets).

Machine Learning Model	Mean Accuracy (%)	SD (%)	Maximum Accuracy (%)	Minimum Accuracy (%)
Logistic regression	85.31	2.46	90.20	82.36
Random forest	85.31	2.18	88.57	82.00
Support vector machine	83.85	1.72	86.75	81.45
Ridge regression classifier	83.83	2.00	86.93	80.73

SD, standard deviation.

TABLE 2. Average processing time with machine learning methods.

Machine Learning Model	Average processing Time (sec)	SD
Logistic regression	16.98	0.62
Random forest	708.43	42.08
Support vector machine	18.28	1.05
Ridge regression classifier	33.77	2.31

SD, standard deviation.

Note: Processing time based on OSX EI Capitan, Processor 2.4 GHz Intel Core i5, Memory 8 GB 1333 MHz DDR3.