Mining HIV Trends in Social Media Data

Patrick Breen
Institute of Bioinformatics
The University of Georgia
Athens, Georgia
pbreen@uga.edu

Shannon Quinn*
Department of Computer Science
University of Georgia
Athens, Georgia
squinn@cs.uga.edu

ABSTRACT

Pre-Exposure Prophylaxis (PrEP) is a recently developed therapy for the prevention of Human Immunodeficiency Virus (HIV) transmission. The treatment is available as a pill named Truvada and is taken once a day by HIV negative individuals for as much as 99% protection when exposed to HIV. While medically effective, PrEP suffers from high drug costs, HIV-related social stigma, and under-informed health providers and patients. Data mining of social media has proven effective in the monitoring of HIV in the US, but no study has investigated PrEP using social media. This paper describes a data mining and machine learning strategy using natural language processing that monitors Twitter social media data to identify PrEP discussion trends. Our results show that we can identify PrEP and HIV discussion dynamics over time, and PrEP-related tweets with positive and negative sentiment. These results can be used by public health professionals to monitor PrEP discussion trends, and identify strategies to improve HIV prevention.

Keywords

Social Media, HIV, Topic Modelling, Document Classification

1. INTRODUCTION

Pre-exposure prophylaxis (PrEP) is an HIV prevention method marketed in the form of a pill named Truvada. Truvada was approved by the Food and Drug Administration (FDA) in 2012 to prevent sexually acquired HIV infection after several positive trials [6, 13] showed that it was safe and effective. Despite minimal side effects, and risk protection of up to 99%, PrEP suffers from uncertain health insurance compensation, the risk of spreading HIV drug resistance and uninformed health providers and patients [8]. Truvada must be taken once a day for full protection. In some individuals, adherence is difficult to maintain [15], leading to loss of Truvada's effectiveness. While public health officials can spread information and monitor the effectiveness of PrEP at the clinic-level, scaled-up data mining on social media data may provide more complete information on PrEP at the national level.

Twitter has been used as a source of data for large scale opinion mining in public health monitoring contexts to predict the spread of influenza [1], predict postpartum depression [4], and examine tobacco use [12]. Recently it has also been used for the study of HIV [18, 17]. These studies of HIV have focused on county-level HIV prevalence prediction, and general HIV discussion monitoring, but they have not focused on PrEP related discussion. Furthermore, existing HIV social media analyses have not taken full advantage of natural language processing (NLP) techniques to discover semantic information in unstructured text [16].

Tweets from twitter consist of short 140 character text messages that may also contain hashtag annotations.

One recently developed tool for NLP, called Word2Vec, is a connectionist method that embeds words as word-vectors in a semantic space that captures substitution-similarity [11]. There are several forms of Word2Vec, though the most popular version, Skip-gram Negative Sampling (SGNS), has been shown to perform well at producing word-vectors that capture important word relationships. This includes most notably word analogies.

Word2Vec is also used as a preprocessing step for additional analyses that start from pre-trained word-vectors. Doc2Vec [7] is a method that produces a document-vector for each document, and each document-level identifier. In the case of a tweet-corpus, document identifiers might include for example document ID, hashtags, and the user who created the document. Each of these document-level identifiers is embedded in a similarity space, allowing one to identify similar tweets, hastags and users.

Latent Dirichlet Allocation [3] (LDA) is used to identify a small set of latent topics present in an unstructured corpus. LDA is a graphical model that generates documents from a set of latent topics. A topic is a probability distribution over words that captures a set of related words. LDA models are often inferred in practice using Bayesian inference via either collapsed Gibbs sampling or Variational Bayesian inference. Inspection of the resulting topics allows one to identify relevant terms and the context in which they occur in the corpus.

Dynamic Topic Modeling [2] (DTM) is an extension of LDA that produces a seried of topic models over time. Briefly,

*Corresponding author.

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

the documents in the corpus are divided into several corpuses that are successive in time. LDA is performed on each corpus, to extract a topic distribution. The posterior topic distribution from time t_n is used as the prior for time t_{n+1}. This lets the DTM model determine a topic model for each time point that is dependent both on the set of tweets from that time point, and on the previous time point's document distribution.

In this paper we answer the acknowledged need to harness large scale data in the battle against HIV [16]. We use the above NLP techniques to extract PrEP-related semantic information from a Twitter corpus dataset. We identify critical PrEP related terms, users, hashtags and tweets. We identify PrEP discussion trends over time, and identify other topics that people who tweet about PrEP also tweet about. Finally we train a sentiment classifier that automatically identifies PrEP related tweets with positive and negative sentiment. Together these results, and these approaches can be used by public health officials to identify the national PrEP discussion and respond to issues as they arise.

2. RESULTS

We sought to determine trends in HIV and PrEP discourse on twitter to inform and coordinate public health efforts aiming to promote PrEP adoption and adherence for at-risk individuals. We collected 624,569 tweets containing at least one of the following words 'HIV', 'AIDS', 'truvada', 'prophylaxis', 'imtesting', 'PrEP' from Twitter's streaming API (TODO: how to cite Twitter streaming API?). The tweets were restricted to English language, and the collection dates spanned from the 47th week of 2015 to the 14th week of 2016. The tweets were cleaned of exotic characters. Before performing topic analysis, we excluded words that were mentioned fewer than 10 times, or more than 0.3 times the number of documents. We also performed Term Frequency Inverse Document Frequency (TF-IDF) normalization.

We excluded tweets that did not originate in contiguous United States time zones. 14204 of the tweets we collected (about 2%) had geolocation coordinates available. We found that tweets were largely concentrated on US and Canadian metro areas (Figure 1). The tweets did not seem to be over represented in any geographical region of the US.

2.1 Word and Document Similarity

The first analysis that we performed sought to identify certain keywords, hashtags, tweets and users, that were discussed in HIV and PrEP related contexts. Word2Vec and the related method, Paragraph2Vec are unsupervised machine learning methods that have performed well at embedding natural language in a semantic vector space. In our analysis, Word2Vec allows us to determine semantically similar words to a query word, while Paragraph2Vec allows us to determine similar tweets, users and hashtags to a query hashtag.

We trained a Word2Vec model, and queried for the top 10 word-vectors related to the term PrEP (Table 1). We found several HIV/AIDS related events, WorldAIDSDay, HLM2016AIDS, ICASA2015, as well as the PrEP drug Truvada, and the term HIV. In addition, we found the acronym ART which refers to Anti-Retroviral Therapy. DoingIt, and OneConversation, are ongoing efforts led by the Center for Disease Control (CDC) to spread awareness and reduce the spread of HIV. NBHAAD is an organization that is committed to increasing awareness for HIV within the Black community. Nancy Reagan, who died in early 2016, was mentioned in conjunction with her efforts to combat HIV in the 1980's. Together these results show us a high-level view of the important components of the national PrEP conversation on Twitter.

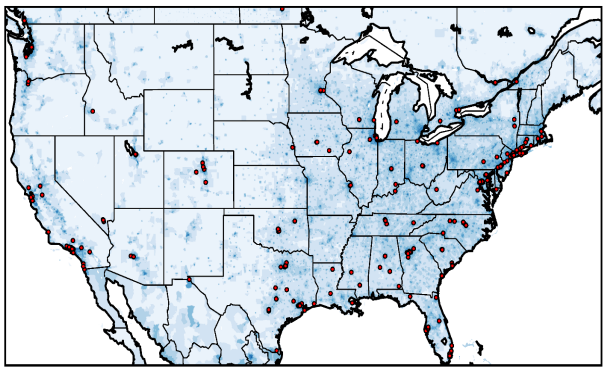

Figure 1: Plot of geolocated tweets.

Table 1: Cosine Similarity to word-vector PrEP

Related Word	Cosine Similarity to PrEP
truvada	0.796666
DoingIt	0.738141
WorldAIDSDay	0.720910
NancyReagan	0.717667
NBHAAD	0.705061
ART	0.704300
HLM2016AIDS	0.698698
ICASA2015	0.693117
HIV	0.692860
OneConversation	0.688427

Doc2Vec allows us to identify the top users, tweets, and hashtags associated with #prep. Note that on Twitter, hashtags are not case sensitive. Querying for the top 10 document-level entities associated with #prep, we again see several PrEP and HIV related hashtags including #hiv, #hivprevention, #truvada, and #whereisprep. We also see a LGBT-related hashtag, #lgbtmedia16, which indicates a distinct awareness of PrEP in the Gay community. This may reflect the known levels of HIV transmission in men who have sex with men [5]. Together the Doc2Vec results show that we can monitor and identify PrEP-related hashtags and tweets.

Interestingly, we also found 3 tweets and 1 user in the top 10 doc2vec results for #prep. Tweet 702179860983189504 has content spreading HIV prevention awareness: "#StoneCold-VideoTODAY if You see this 13 symptoms. Do HIV Test Immediately. Must Read". Conversely tweet 708519265540907010 has content that calls into doubt the usefulness of PrEP:

Table 2: Cosine Similarity to doc-vector #PrEP

Related Hashtag/Tweet	Cosine Similarity to #PrEP
#lgbtmedia16	0.739128
#hiv	0.727602
#whereisprep	0.707165
#truvada	0.696113
#hivprevention	0.636068
tweet-702179860983189504	0.630055
user-711275699529764864	0.629254
tweet-708519265540907010	0.628778
tweet-712032637024653313	0.628646
#harrogatehour	0.628547

"Checkout why PrEP is hurting the cause & #JoinTheConversation #LGBTQIA". User 711275699529764864 appears to be a twitter spam bot with no obvious connection to PrEP. Thus with this method we demonstrate a way to identify and monitor the most viral tweets related to PrEP. As the discussion changes, public health professionals can use this information to quickly identify the most relevant viral sentiment in the online PrEP conversation.

Finally we wanted to visualize the relative similarities of several PrEP-related keywords in a low dimensional space. We took keywords that we had identified in our PrEP related queries, along with other HIV-prevention related terms, and visualised their word-vectors in 2 dimensions using tSNE [14] (Figure 2.

We several trends. Notably the pharmaceutical based HIV therapies all cluster together (ART, PrEP, truvada) and the AIDS awareness events cluster together (WorldAIDSDay, NBHAAD, ICASA2015). Words that are related to HIV discussion, but also used in other contexts (undetectable, testing, awareness) are further away from the HIV/AIDS word-clusters. These results provide another mechanism for researchers to visualize and identify relevant trends in PrEP-related keywords.

2.2 Time Domain

Next we sought to identify some temporal trends in PrEP related trends. We used Dynamic Topic Modelling (DTM) to identify how certain topics change over time. We specified 10 topics and used each week's worth of tweets, over a 30 week period from the 47th week of 2015 to the 14th week of 2016, as our time points.

We identified two topics that showed relevant trends. In topic 5 we see that the keyword 'PrEP' increases over time, while 'prevention', 'drug' and 'risk' remain constant (Figure 3). The term 'pill', also from topic 5, declines overtime. These dynamics may indicate that PrEP discussion is becoming more prevalent. This growing trend would be consistent with the fact that while PrEP is still not widely known about among patients and healthcare providers, information about PrEP is slowly building into national awareness. A study in New York City in 2011 indicated that only 36% of high risk individuals were aware of PrEP [10].

Other HIV prevention related words such as 'pill', 'prevention' and 'drug' serve as negative controls. They are related to PrEP, but also used in other medical contexts. The fact that they are not increasing, shows us that the increase in PrEP discussion is PrEP-specific.

However, it is hard to tell from these data whether the

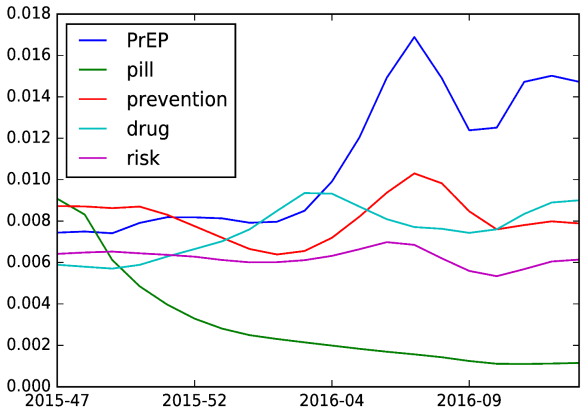

Figure 3: DTM topic 5 (PrEP related topic) word prevalence over time. Date is YYYY-WW.

increased level of PrEP discussion is leading to increased levels of informed patients, medical providers, and adherence. It is possible that stigma, and misinformation is leading to greater levels of PrEP discussion on twitter. We will get more specific, granular understanding of the PrEP discourse in our sentiment analysis section below (see section "Sentiment Classification").

We found at least one other DTM topic that showed interesting behaviour. We found that topic 4 captured several keywords related to World AIDS Day (Figure 4). We can see that "WorldAIDSDay", and "Can" peak in the 47th week of 2015 and then decline into 2016. This correlates well with the actual date of World AIDS Day, December 1st. Furthermore, while December 1st is World AIDS day, the whole month of December is AIDS Awareness Month. We can clearly see the words 'raise' and 'awareness' peak later and last longer than the word 'WorldAIDSDay' implying that it is correlated to the whole month of December. While our PrEP investigation isn't specifically interested in World AIDS Day, or AIDS Awareness Month, this observation validates our ability to accurately identify temporal events using DTM.

Together, the DTM results demonstrate our ability to extract relevant HIV and PrEP related information from Twitter that accurately captures time-dependant fluctuations. Public health professionals should be able to monitor these temporal trends to determine the relative interest in PrEP, and other HIV related keywords as they are discussed over time.

2.3 User Timeline Analysis

We wanted to identify what Twitter users that mentioned PrEP were discussing in their other tweets. We identified users that were most similar to PrEP according to our Paragraph2Vec results, and downloaded their most recent 3000 tweets. We took the top 500 users that had at least 200 words in the combined tweets of their tweet history. For each user, we concatenated all of their tweets, and performed LDA topic modelling on the resulting set of user-timeline documents (Figure 5).

We performed LDA on the users' timelines using the pyL-

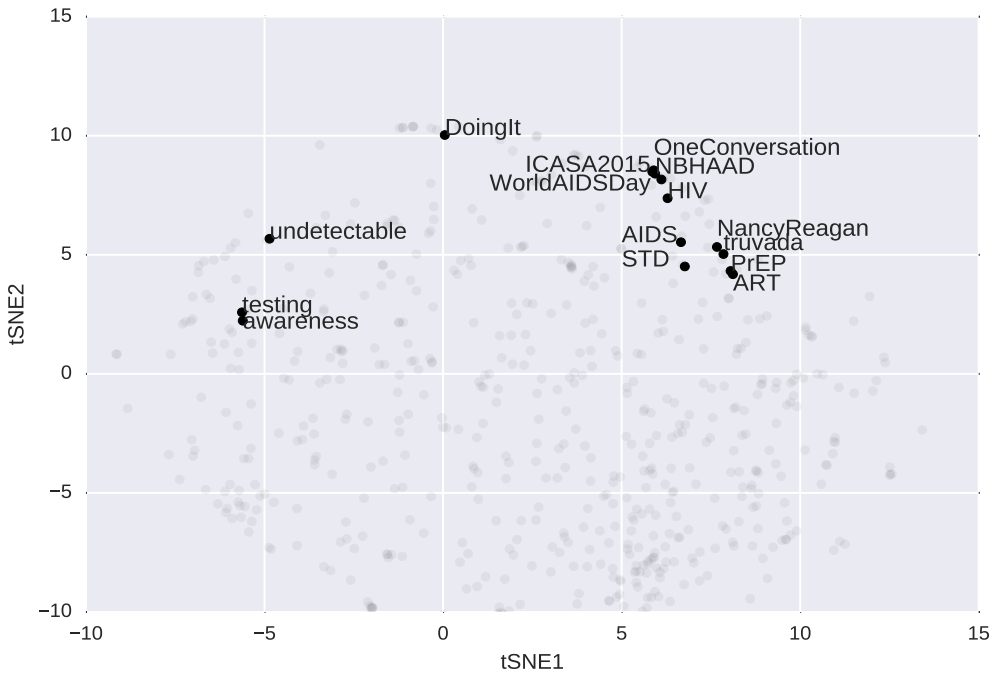

Figure 2: tSNE plot of relevant word-vectors

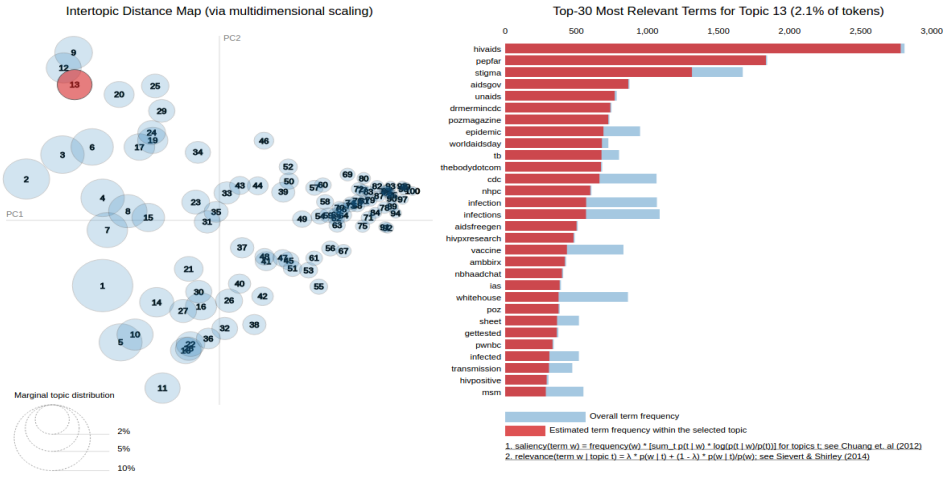

Figure 5: Clustermap of users vs. topics for the top 400 users related to PrEP.

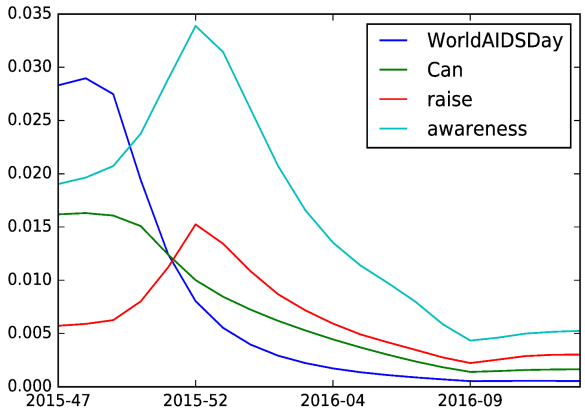

Figure 4: DTM topic 4 (WorldAIDSDay related topic) word prevalence over time. Date is YYYY-WW.

DAvis python package from graphlab [9]. We specified 100 topics, and found that only one of them was related to HIV (topic number13, see Figure 5) representing about 2% of the marginal topic distribution. Topic 13 contained top terms "hivaids", "pepfar", "stigma" and "aidsgov". PEPFAR is a governmental organization, The United States President's Emergency Plan for AIDS Relief, that is focused on combating the spread of AIDS internationally. Interestingly, we don't see PrEP in the top 30 terms in topic 13. This indicates that among twitter users that talk about PrEP, HIV is a small part of their discussion, and PrEP an even smaller part of their discussion.

We investigated some of the other major topics from users' timelines, and found discussions of other STD's (topics 20, 29 and 17) and other health related terms (topics 12, and 25). The other health and STD-related topics clustered closely with topic 13 in Principle Component (PC) space. Some of the STD-related topics included terms related to LGBTQ, including the social networking platform Grindr in topic 17. Other terms related to STD's include prevention terms such as "CDC", "condoms", and "gettested". Topic 19 is nearby these STD-related topics in PC space, and contains terms related to healthcare and political issues such as "Obamacare" and "ACA" (Affordable Care Act). This connection between HIV and other STD prevention, and political discourse may be relevant in PrEP-based HIV prevention efforts, considering there is in some cases no clear precedent for how preventative therapies like PrEP are covered by health insurance [8].

One of the top words from the HIV-related topic, topic 13, was the term "stigma", the term "endstigma" was also found in topic 17. Previous studies have shown a variety of stigmas associated and HIV have hampered prevention efforts [8]. Our observation of stigma related terms corroborates that there is some discussion of stigma in the context of HIV on Twitter. Public health professionals may be able use the prevalence of the term "stigma" as a way to monitor the efficacy of efforts to end HIV related stigmas.

2.4 Sentiment Classification

Previously, we used analyses to summarize the whole Twitter corpus to produce high level trends. For our final analysis, we sought to get a deeper understanding of the data at the individual tweet level. Thus we trained a classifier to classify the sentiment of HIV and PrEP related tweets either positive or negative. This classifier would allow public health professionals to quickly identify positive and negative PrEP related tweets to guide HIV prevention efforts. We obtained a set of 1.6 million tweets with sentiment labels, either positive or negative from Sanders Analytics (http://www.sananalytics.com/lab sentiment/). Then we trained a simple logistic regression classifier on 1.2 million paragraph-vectors from the sentiment dataset. We found that our classifier had an accuracy of 69% using a portion of the sentiment tweets not used in training, as a validation set.

We chose to use a relatively simple classifier model (logistic regression) and stop training at 20 iterations of stochastic gradient decent over the corpus, because we wanted to prevent the possibility that we overtrained on our training data. This was especially important because our training and testing data, while both sets of tweets, were separate datasets. We then used this classifier to classify our PrEP related tweets into positive or negative sentiment labels. We identified the most positive, and most negative tweets, by log probability, on our full dataset, and on tweets that specifically mention either PrEP or Truvada. We provided the text from the top 3 positive and negative tweets from each dataset (Table 3, Table 4).

While the sentiment classification doesn't have perfect accuracy, we can see positive tweets disseminating information about the efficacy and benefits of PrEP related preventions (Table 3). The positive tweets indicate that PrEP public health efforts have had some effect disseminating PrEP information on Twitter.

The negative tweets may be more important than the positive tweets to guide future public health policy corrections. The negative tweets contain concerns that Truvada may not block HIV transmission in all cases (Table 4). There also seems to be concern that use of Truvada may increase the transmission of non-HIV STDs such as Hepatitis C because some PrEP users stop using condoms. This may indicate that public health professionals need to stress that PrEP users should not stop wearing condoms when they spread PrEP related information.

Another negative tweet contains questions over weather Truvada is available as a generic drug. This may indicate that in some cases, patients are aware of PrEP, but do not have access to it, either because it is not available through their healthcare provider, or because it is not affordable.

Finally we see some contention over national political leadership in the effort to prevent HIV. While political discussions can often get heated, especially on Twitter, the contention can harm concerted efforts to provide consent governmental leadership in HIV prevention. Public health officials may consider stressing that unity in health policy at the federal level is important to protect the population from the spread of HIV.

3. CONCLUSIONS

- Moving towards HIV outbreak prediction.
- Automatically identifying PrEP adoption/adherence sentiment in US population.
- A lot of information on twitter is from bots, they only

Table 3: Positive Sentiment Tweets.

Category	Text
General	"RT TOPublicHealth The Works provides testing for HIV anonymous & rapid test available . Call 416-392-0520 for more info"
General	"RT FCAA ejaforg announced 5.4 million in grants to support orgs addressing #HIV in new & innovative ways!"
General	"RT HillaryClinton A note on the fight against HIV and AIDS and the people who really started the conversation."
PrEP specific	"He won't use condoms because intimacy means more than his health. but he's discovered PrEP. thank goodness."
PrEP specific	"PrEP Queensland Aids Council, #HIV Foundation, Queensland."
PrEP specific	"RT JDatTheBody At the core of our programs is belief that young ppl can succeed in take PrEP for HIV prevention. #NHPC2015"
Truvada specific	"RT CDC_HIVAIDS Expanding testing, treatment, & #PrEP could prevent up to 185k new #HIV infections"
Truvada specific	"Another reason 4 #Ireland & #UK 2 immediately approve #truvada & #PrEP 2 stop #HIV infections . arleavitt AodhanORiordain MerchantsQuayIR"
Truvada specific	"RT EvanJPeterson For #worldAIDSday my early #PrEP arcticle in strangerslog, art by leviathanleague #hiv #truvada #truvadawhore"

Table 4: Negative Sentiment Tweets.

Category	Text
General	"Also, how fucking vile of Hillary to say. Reagan did fucking NOTHING during the AIDS epidemic until it was too late. What a stupid old hag."
General	"I wonder why he beat her ass when she was tryna leave like she wasn't gone be running back when she found out she had HIV & nobody want her"
General	"Aaannd. Hillary Clinton breathes a sigh of relief that Twitter has left its outrage of her AIDS comments behind to tend to Drumpf debacle."
PrEP specific	"RT gaston_croupier #Truvada patent's not expired yet but it is sold online as a generic drug? There's something rotten in internet #PrEP h"
PrEP specific	"Equality_MI Syph & Hep C have gone up 550% in Gay Men bc many feel tht bc they're on PrEP, they don't need condoms. HIV isn't the only STI."
PrEP specific	"Xaviom8 in interviews he says he was adherent. strain was highly resistant, and Truvada wouldn't have blocked it anyways. PrEP didn't fail."
Truvada specific	"not surprised at all that someone got HIV on truvada. people get pregnant on birth control. tomato-condoms are still important-tomahto"
Truvada specific	"Now reading that truvada does not protect against certain strains of the HIV virus. Yet people want to take that risk.."
Truvada specific	"I think I have conjunctivitis unless truvada cured it overnight cuz im not feeling as horrible today as last night"

cost $ per bot. This along with the character limit and slang makes twitter hard to use, but we've shown that it can be used. In fact twitter bots are slowly making twitter unusable because free advertising and marketing. Maybe in a future project we make twitter bots to spread information about PrEP.

4. ACKNOWLEDGMENTS

Acknowledgements (optional) go here.

5. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM, 2013.

[5] C. for Disease Control, P. (CDC, et al. Hiv risk, prevention, and testing behaviors. national hiv behavioral surveillance system: men who have sex with men, 20 us cities, 2011. in: Hiv surveillance special report 8. *HIV surveillance special report*, 8, 2014.

[6] R. M. Grant, J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, P. Goicochea, M. Casapía, J. V. Guanira-Carranza, M. E. Ramirez-Cardich, et al. Preexposure chemoprophylaxis for hiv prevention in men who have sex with men. *New England Journal of Medicine*, 363(27):2587–2599, 2010.

[7] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[8] A. Liu, S. Cohen, S. Follansbee, D. Cohan, S. Weber, D. Sachdev, and S. Buchbinder. Early experiences implementing pre-exposure prophylaxis (prep) for hiv prevention in san francisco. *PLoS Med*, 11(3):e1001613, 2014.

[9] Y. Low, J. E. Gonzalez, A. Kyrola, D. Bickson, C. E. Guestrin, and J. Hellerstein. Graphlab: A new framework for parallel machine learning. *arXiv preprint arXiv:1408.2041*, 2014.

[10] S. A. Mehta, R. Silvera, K. Bernstein, R. S. Holzman, J. A. Aberg, and D. C. Daskalakis. Awareness of post-exposure hiv prophylaxis in high-risk men who have sex with men in new york city. *Sexually transmitted infections*, pages sti–2010, 2011.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8):e174, 2013.

[13] M. C. Thigpen, P. M. Kebaabetswe, L. A. Paxton, D. K. Smith, C. E. Rose, T. M. Segolodi, F. L. Henderson, S. R. Pathak, F. A. Soud, K. L. Chillag, et al. Antiretroviral preexposure prophylaxis for heterosexual hiv transmission in botswana. *New England Journal of Medicine*, 367(5):423–434, 2012.

[14] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[15] A. Van der Straten, L. Van Damme, J. E. Haberer, and D. R. Bangsberg. Unraveling the divergent results of pre-exposure prophylaxis trials for hiv prevention. *Aids*, 26(7):F13–F19, 2012.

[16] S. D. Young. A âĂIJbig dataâĂİ approach to hiv epidemiology and prevention. *Preventive medicine*, 70:17–18, 2015.

[17] S. D. Young and D. Jaganath. Online social networking for hiv education and prevention: a mixed methods analysis. *Sexually transmitted diseases*, 40(2), 2013.

[18] S. D. Young, C. Rivers, and B. Lewis. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*, 63:112–115, 2014.