Mining Pre-Exposure Prophylaxis Trends in Social Media

***Abstract*—Pre-Exposure Prophylaxis (PrEP) is a groundbreaking biomedical approach to curbing the transmission of Human Immunodeficiency Virus (HIV). Truvada, the most common form of PrEP, is a combination of tenofovir and emtricitabine and is a once-daily oral mediation taken by HIV-seronegative persons at elevated risk for HIV infection. When taken reliably every day, PrEP can reduce one's risk for HIV infection by as much as 99%. While highly efficacious, PrEP is expensive, somewhat stigmatized, and many health care providers remain uninformed about its benefits. Data mining of social media can monitor the spread of HIV in the United States, but no study has investigated PrEP use and sentiment via social media. This paper describes a data mining and machine learning strategy using natural language processing (NLP) that monitors Twitter social media data to identify PrEP discussion trends. Results showed that we can identify PrEP and HIV discussion dynamics over time, and assign PrEP-related tweets positive or negative sentiment. Results can enable public health professionals to monitor PrEP discussion trends and identify strategies to improve HIV prevention via PrEP.**

I. Introduction

Pre-exposure prophylaxis (PrEP), a treatment for the prevention of Human Immunodeficiency Virus (HIV), is marketed under the trade name Truvada. Truvada, a combination of tenofovir and emtricitabine, was approved by the Food and Drug Administration (FDA) in 2012 to prevent the transmission of HIV to seronegative persons at elevated risk for HIV infection. Several randomized clinical trials showed that PrEP was safe and could reduce one's risk for HIV infection by up to 99% if taken reliably [1], [2]. Despite its minimal side effects and high level of risk protection, PrEP suffers from uncertain health insurance compensation, the risk of producing drug resistant strains, and uninformed health providers and patients [3]. Truvada must be taken once a day for maximum protection. Consistent adherence to PrEP is difficult for some individuals, leading to lower protection rates [4]. While public health officials can disseminate information about and monitor the efficacy of PrEP at the clinic-level, scaled-up data mining on social media data may provide more complete information on facilitators and barriers to PrEP use at the national level.

Twitter, a popular microblogging service, is the social media platform we use in this study. "Tweets" from Twitter consist of short 140 character text messages that may also contain hashtag annotations. 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 [5], predict postpartum depression [6], and examine tobacco use [7]. Recently it has also been used for the study of HIV [8], [9]. 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 [10].

In this paper we respond to the acknowledged need to harness large scale data in the battle against HIV [10]. We make extensive use of 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 co-occur with who tweet about PrEP. 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 trends in the national PrEP discussion and respond to public health issues as they arise.

II. Methods and Data

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 [12] is one such method that uses pre-trained word-vectors to produce a document-vector for each document-level attribute. In the case of a tweet-corpus, document identifiers might include tweet ID, hashtags, and the user who created the tweet. Each of these document-level identifiers is embedded in a similarity space, allowing one to identify similar tweets, hastags, and users.

Latent Dirichlet Allocation [13] (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 either via collapsed Gibbs sampling or variational Bayesian methods. Inspection of the resulting topics allows one to identify relevant terms and the context in which they occur in the corpus.

Dynamic Topic Modeling [14] (DTM) is an extension of LDA that produces a series of topic models over time. Briefly,

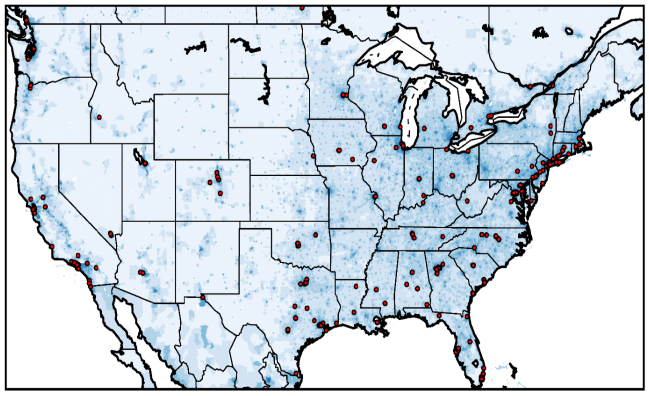

Fig. 1. Plot of geolocated tweets.

the documents in the corpus are divided into several corpora 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 topic distribution.

We collected 624,569 tweets containing at least one of the following words 'HIV', 'AIDS', 'truvada', 'prophylaxis', 'imtesting', or 'PrEP' from Twitter's 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 (non-punctuation, non-alphanumeric 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. A subset of 14,204 tweets (about 2%) had geolocation coordinates available. We found that these tweets were largely concentrated on US and Canadian metro areas (Fig. 1). The tweets did not seem to be over represented in any geographical region of the US.

III. 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.

A. 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, doc2vec, 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 doc2vec allows us to determine similar tweets, users and hashtags to a query hashtag.

TABLE I
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

We trained a word2vec model, and queried for the top 10 word-vectors related to the term PrEP (Table I). 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 over the collection period.

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 an LGBT-related hashtag, #lgbtmedia16, which indicates a distinct awareness of PrEP in the Gay community. This may reflect the elevated levels of HIV transmission in men who have sex with men [15]. 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 (Table II). Tweet 702179860983189504 has content spreading HIV prevention awareness: "#StoneColdVideoTODAY 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: "Checkout why PrEP is hurting the cause & #JoinTheConversation #LGBTQIA".

We investigated the blog article linked to tweet 708519265540907010 [16]. The article, authored by Steven Banning, an HIV positive individual, expresses concerns about how new HIV treatments, like PrEP are being adopted by the HIV community. Steven notes that treatments for HIV positive individuals, and treatments to prevent HIV such as PrEP, have made HIV a much more treatable disease than it was in the 1980's. These medical advances may have led in part to some unintended medical and social consequences.

TABLE II
COSINE SIMILARITY TO DOCUMENT-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

Steven notes that the rise of drug-resistant HIV strains has increased, in part due to HIV patients not adhering to to consistent treatment plans. He also points out the recent social trend of "bug chasers," individuals who are actively seeking to acquire and spread HIV strains. Steven Banning suggests that the widespread availability of effective HIV medications, may have desensitized a healthy sense of fear of the disease in some individuals.

User 711275699529764864 appears to be a Twitter spam bot with no obvious connection to PrEP. Unfortunately, bots are commonly used on Twitter for advertising and marketing purposes, and sometimes hinder pertinent information retrieval. Thus while we observe some noise, our results demonstrate a method to quickly 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 visualized their word-vectors in 2 dimensions using t-distributed Stochastic Neighborhood Embedding(tSNE) [17] (Fig. 2).

We identified 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 demonstrate another mechanism for researchers to visualize and identify relevant trends in PrEP-related keywords.

B. Time Domain

Next we sought to identify some temporal trends in PrEP related trends. We used Dynamic Topic Modeling (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,

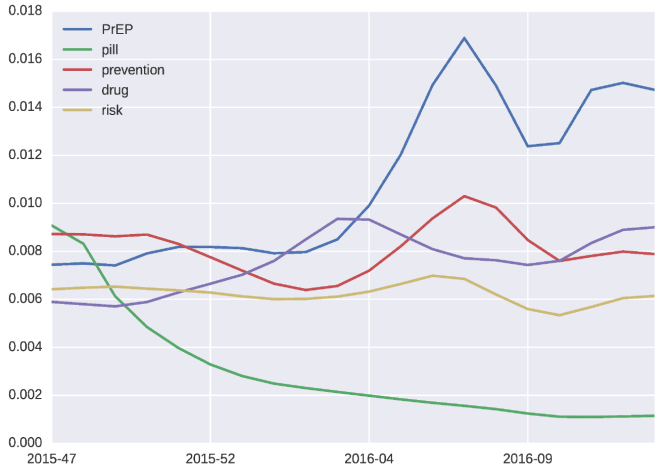

Fig. 3. DTM topic 5 (PrEP related topic) word prevalence over time. Date is YYYY-WW.

while 'prevention', 'drug' and 'risk' remain constant (Fig. 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 entering into national awareness. A study in New York City in 2011 indicated that only 36% of high risk individuals were aware of PrEP [18].

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 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 behavior. We observed that topic 4 captured several keywords related to World AIDS Day (Fig. 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' indicating that these words are 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 serves as a control to validate our ability to accurately identify temporal

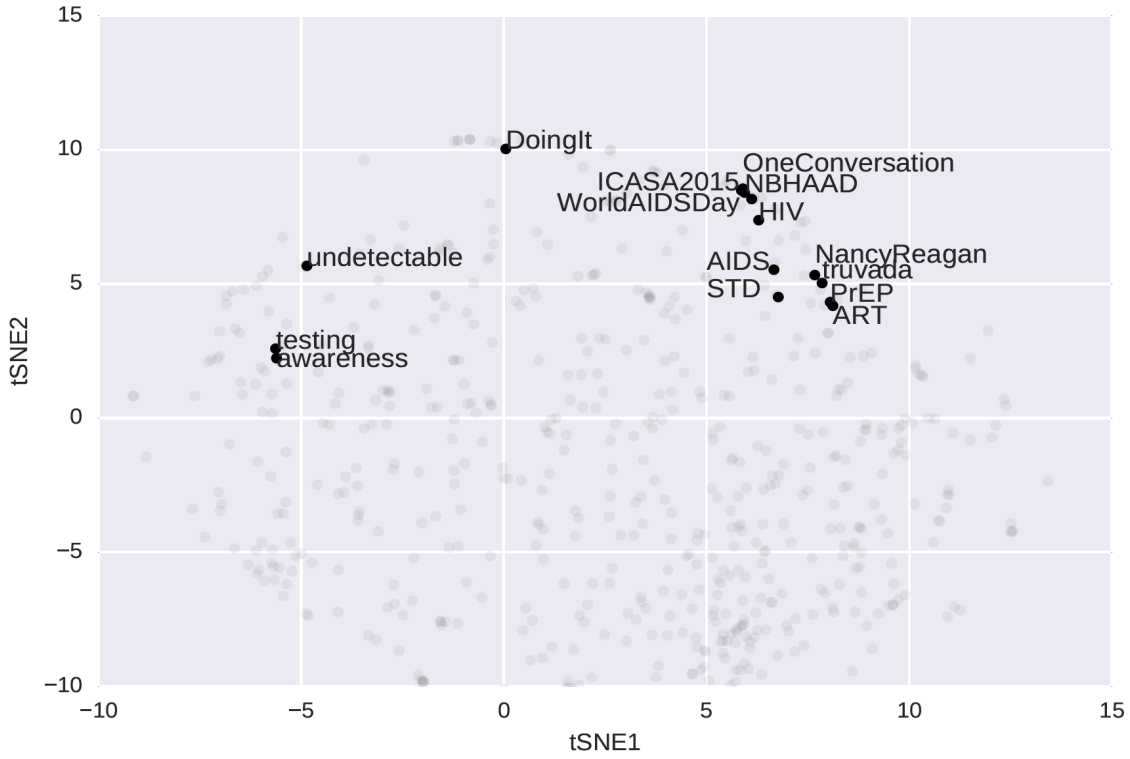

Fig. 2. tSNE plot of relevant word-vectors

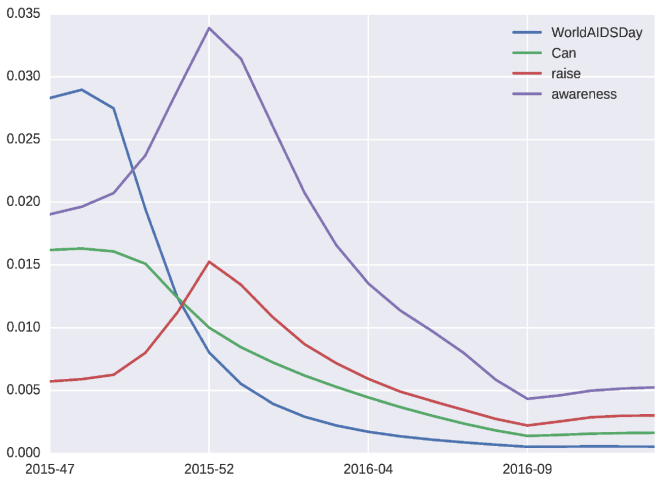

Fig. 4. DTM topic 4 (WorldAIDSDay related topic) word prevalence over time. Date is YYYY-WW.

events using DTM.

Together, the DTM results demonstrate our ability to extract relevant HIV and PrEP related information from Twitter that accurately captures time-dependent 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.

C. 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 using our doc2vec results, then downloaded their recent tweet history, up to their last 3000 tweets. We selected the top 500 users that had at least 200 total words in the combined tweets of their tweet history. For each user, we concatenated their timeline of tweets, and performed LDA topic modeling on the resulting set of user-timeline documents (Fig. 5).

We performed LDA on the users' timelines using the pyLDAvis python package from GraphLab [19]. We specified 100 topics, and found that only one of them was related to HIV (topic number 13, 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,

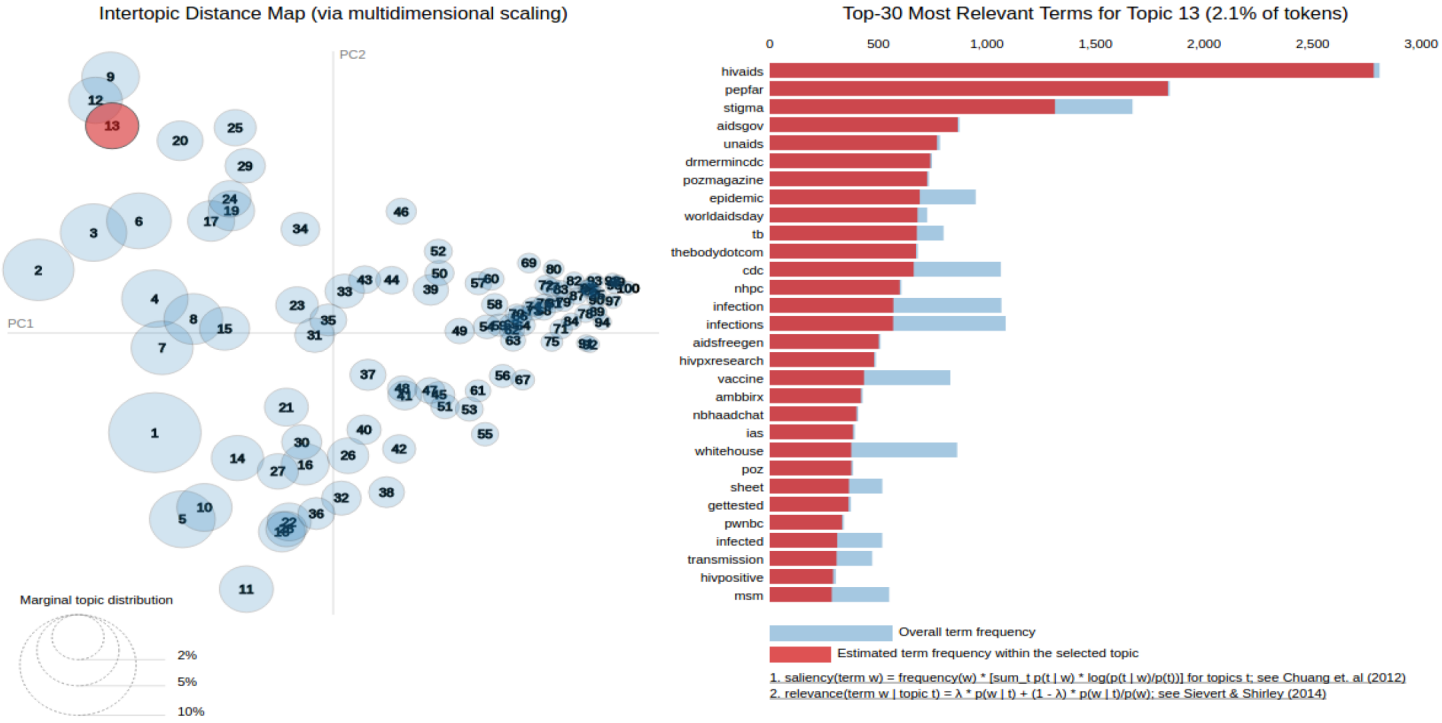

Fig. 5. LDA topic modeling for the top 500 users related to PrEP.

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, other STD prevention topics, 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 [3].

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 [3]. 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 "endstigma" as a way to monitor the efficacy of efforts to end HIV related stigmas.

D. Sentiment Classification

Previously, we used analyses to summarize the whole Twitter corpus to extract 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 to be 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 binary sentiment labels from Sanders Analytics [20]. 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 epochs 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 used this trained classifier to classify our PrEP related tweets into positive or negative sentiment labels. We identified the most positive, and the most negative tweets, by log probability, on our full dataset, and positive and negative tweets that specifically mention either PrEP or Truvada. We provide the text from the top three positive and negative tweets from each of the three datasets (Table III, Table IV).

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

TABLE III
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 IV
NEGATIVE SENTIMENT TWEETS.

Category	Text
General	"Also, how f***ing vile of Hillary to say. Reagan did f***ing NOTHING during the AIDS epidemic until it was too late. What a stupid old hag."
General	"I wonder why he beat her a** 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"

HIV transmission in all cases (Table IV). 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 continue to wear condoms for full protection against other diseases when disseminating PrEP-related health information.

Another negative tweet questions whether Truvada is available as a generic drug. PrEP is currently only available through Gilead's Truvada, though a generic may be available as soon as 2017 [21]. Current patients must deal with uncertain drug prices, which vary from $14,000 to $70 [21], and have in some cases, uncertain medical insurance reimbursement status. This tweet highlights both the difficulty of acquiring affordable PrEP medication, and also a level of conflicting information circulating on social media.

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 consistent 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.

The last tweet in the positive tweets (Table III), referencing the hashtag "#truvadawhore", links to a blog article written by Evan Peterson titled "The Case for PrEP, or How I Learned to Stop Worrying and Love HIV-Positive Guys" [22]. Evan is an HIV-negative individual who adheres to a daily PrEP regimen to protect himself from HIV. Evan finds that for himself, the side effects and inconveniences associated with taking daily Truvada are worth the protection, noting that he also practices safe sex, and gets tested regularly. Evan addresses the hashtag #truvadawhore, which originated with a Huffington Post article titled "Truvada Whores?" [23] which expressed concern toward the increased sexual risk taking of some Truvada patients. Evan provides a rebuttal, noting that many HIV-negative individuals on Truvada, including himself and

people he knows, are responsible and cautious about the risks they are taking. He explains that taking daily Truvada may actually be an indication that individuals are more educated and responsible with respect to their sexual health. Further large scale quantitative measurements, including large scale medical surveys, will be needed to determine whether or not taking Truvada correlates positively or negatively with risk-taking behaviors.

IV. Conclusions

In this article we use a variety of NLP techniques to analyze a tweet corpus in order to monitor the national social media discussion for HIV prevention. We identified several notable trends. We found that PrEP discussion activity is increasing on Twitter, and also that people who mention PrEP also talk about general health, STDs, stigma, and politics.

By picking out the tweets most relevant to PrEP, and classifying these tweets as positive or negative, we quickly identified core concepts important to the online PrEP discussion. We identified two important arguments for and against PrEP. Blogger Steven Banning hypothesizes that PrEP adoption could lead to unintended consequences and unhealthy risk seeking behavior. The contrary hypothesis, described by blogger Evan Peterson, suggests that PrEP adherence is a product of healthy, responsible, at-risk individuals who are gaining an effective form of HIV protection. These opposing arguments were identified automatically using large scale data mining of social media. Further investigation and data analyses of social media data together with other health data may be able to quantify PrEP and Truvada's effectiveness in the ongoing effort to prevent HIV infection, as well as highlight possible strategies to maximize PrEP effectiveness.

Future analyses may also take further advantage of the rich Twitter API. Rather than passively collecting data, intelligent chatbots could be deployed to interact with HIV and PrEP-tweeters. This would enable researchers to conduct large scale social-media-based health surveys and directly observe the dynamics of the discussion.

Acknowledgments

The authors would like to thank many individuals for their consultations in context with the development of this study and its ongoing work.

References

[1] 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*, vol. 363, no. 27, pp. 2587–2599, 2010.

[2] 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*, vol. 367, no. 5, pp. 423–434, 2012.

[3] 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*, vol. 11, no. 3, p. e1001613, 2014.

[4] 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*, vol. 26, no. 7, pp. F13–F19, 2012.

[5] 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*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[6] 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*. ACM, 2013, pp. 3267–3276.

[7] 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*, vol. 15, no. 8, p. e174, 2013.

[8] 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*, vol. 63, pp. 112–115, 2014.

[9] S. D. Young and D. Jaganath, "Online social networking for hiv education and prevention: a mixed methods analysis," *Sexually transmitted diseases*, vol. 40, no. 2, 2013.

[10] S. D. Young, "A big data approach to hiv epidemiology and prevention," *Preventive medicine*, vol. 70, pp. 17–18, 2015.

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

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

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

[15] 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*, vol. 8, 2014.

[16] B. Steven, "why prep is hurting the cause," http://queer-voices.com/2015/08/why-prep-is-hurting-the-cause/, 2015.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[18] 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*, pp. sti–2010, 2011.

[19] 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.

[20] "Twitter sentiment corpus," http://www.sananalytics.com/lab/twitter-sentiment/, accessed: 2016-03-14.

[21] G. Ethan, "When might generic truvada be available for prep in the us?" http://www.ethantheblog.info/?p=50, 2015.

[22] P. Evan, "The case for prep, or how i learned to stop worrying and love hiv-postive guys," http://www.thestranger.com/seattle/the-case-for-prep-or-how-i-learned-to-stop-worrying-and-love-hiv-positive-guys/Content?oid=20991643, 2014.

[23] D. David, "Truvada whores?" http://www.huffingtonpost.com/david-duran/truvada-whores_b_2113588.html, 2012.