Mining HIV Trends in Social Media Data

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

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

ABSTRACT

Here is the abstract.

Keywords

Social Media, Topic Modelling, Document Embedding

1. INTRODUCTION

Introduce and literature review of: 1) PrEP, HIV, Twitter
2) Word2Vec, Doc2Vec
3) Dynamic Topic models, Latent Dirichlet allocation

2. RESULTS

1.2 Million tweets related to 'hiv', 'aids', 'truvada', 'prophylaxis', 'imtesting', 'sexwork', 'gay', 'PrEP', were collected 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 7th week of 2016. The tweets were lowercased, and cleaned of exotic characters.

Then a variety of analyses were run, each with their own additional preprocessing. These analyses sought to determine what things Twitter users were saying about HIV and PrEP, ultimately to determine how to coordinate public health efforts to promote PrEP adoption and adherence for at-risk individuals.

While we excluded non-English tweets, we found that the subset of tweets that had geolocation data available came from around the world. Tweets were concentrated on English speaking areas with high population (Figure 1).

2.1 Word and Document Similarity

The first analyses that we performed sought to identify certain keywords, and hashtags that users were users were mentioning in PrEP related contexts. Word2Vec and the refinement, Paragraph2Vec are unsupervised machine learning methods that have performed well at embedding natural language in a semantic vector space.

*Corresponding author.

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

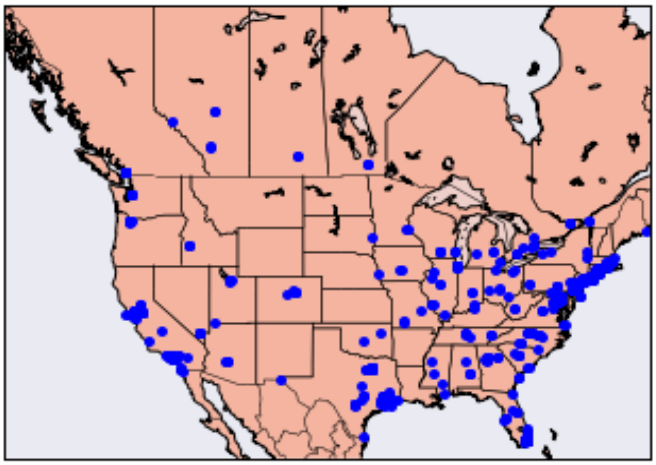

Figure 1: Plot of geolocated tweets.

We queried for the top 10 words that had the highest semantic similarity with the word 'truvada' (Table 1). We found many HIV related words, with the specific word 'prep' as the most similar.

Next we preformed Paragraph clustering. In this application, a paragraph is equivalent to the text of a tweet. In this method, we identify the similarity between tweets, and tweet-level attributes such as hashtags. Querying for the Paragraph-Vectors with high similarity to '#truvada' identifies '#prep' as the top related hashtag (Table 2). Furthermore, we see hashtags related to prevention, and HIV in the results.

The numeric entries in Table 2 correspond to tweet ID numbers. The top tweet associated with '#truvada' has text: "#girlsbelike if you see this 13 symptoms. do hiv test immediately. please read". This tweet is an HIV symptom related news update, which is representative of a large number of HIV related tweets on Twitter.

The combined results from Word2Vec and Paragraph2Vec demonstrate that we can scan through the Twitter corpus and find words, hashtags, and tweets that are semantically related to our PrEP query terms. When HIV public health researchers identify an important keyword, or hashtag, they can quickly use this method to identify related keywords and specific relevant tweets.

2.2 Time Domain

Table 1: Cosine Similarity to 'truvada'

Related Word	Cosine Similarity to 'truvada'
prep	0.836507
charliesheen	0.774535
worldaidsday	0.759242
hivtestweek	0.797361
hiv	0.744817
hivaids	0.730966
aidsfreefuture	0.721897
icasa2015	0.713624
benegative	0.709129
martinshkreli	0.702565

Table 2: Cosine Similarity to '#truvada'

Related Hashtag/Tweet	Cosine Similarity to '#truvada'
#prep	0.740855
#hiv	0.646245
685501883448963072	0.582529
#prevention	0.0.578786
#potus	0.561461
688733331081527296	0.558048
#[Japanese "email friend"]	0.557464
#[Japanese "sex friend"]	0.556990
#bbbh	0.547450
#cure	0.545497

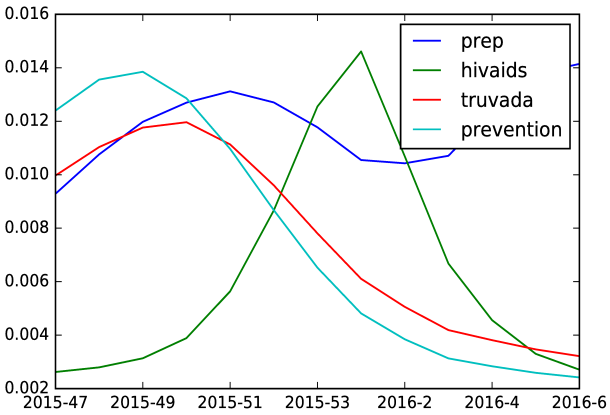

Figure 2: DTM topic 3 word prevalence over time. Date is YYYY-WW.

Next we sought to identify some temporal trends in PrEP related trends. We used Dynamic Topic Modelling to identify how certain topics change over time. We specified 10 topics and plotted representative words from two of the topics.

In topic 3 we see that the keyword 'prep' remains relatively constant, while 'prevention' and 'truvada' decline over time. 'hivaids' increases and then decreases in magnitude over the course of the time series (Figure 2). These dynamics may indicate short term events in hivaids and PrEP-related conversation on twitter over the time period studied.

We found that topic 5 captured several keywords related to World AIDS Day (Figure 3). We can see that all of these terms 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. While we aren't specifically interested in World AIDS Day to inform our understanding of PrEP discussion, this observation validates our ability to identify temporal events using DTM.

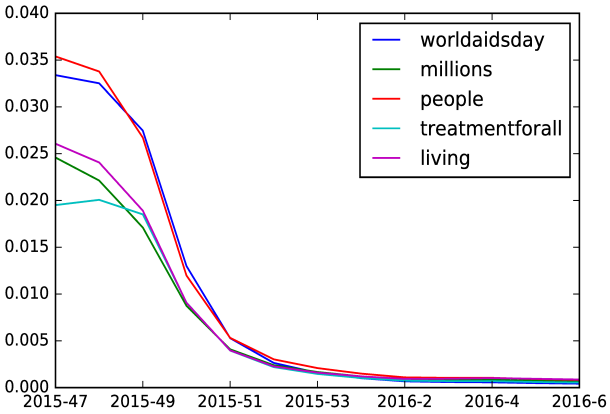

Figure 3: DTM topic 5 word prevalence over time. Date is YYYY-WW.

2.3 User Timeline Analysis

Using a Paragraph2Vec analysis, we queried for the top 500 users related to '#prep' who had at least 200 words in the combined set of tweets in their timeline, up to 3000 tweets. The timelines of these users were concatenated into 500 large timeline-documents. We then performed LDA on the timeline documents and identified words associated with 10 topics (Table 3). We can see some political, news and social media related keywords and colloquialisms in the top 5 words per topic.

A heatmap of Topics vs Users shows the distribution of topic assignment weighting over the 500 users (Figure 4). We can see that many of the users are mentioning things

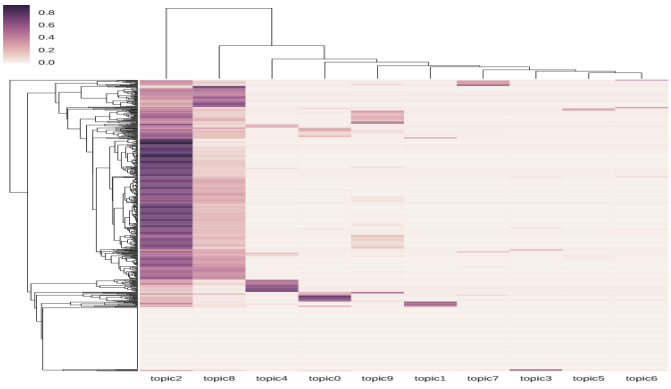

Figure 4: Heatmap of Topics vs. Users.

assigned to topics 2 and 8. Topic 2 appears to be related to certain geographic regions such as 'Nigeria' and 'Syria' that may be especially suffering from HIV. Topic 8 contains words such as 'giveaway' and 'startups' indicating that it may be related to social and economic coordination to combat HIV.

2.4 Sentiment Classification

Lastly we sought to create a classification scheme to classify the sentiment of tweets either positive or negative. This classifier would allow us to quickly identify positive and negative PrEP related tweets to guide public health efforts. We first performed Paragraph2Vec on both our PrEP related tweet corpus and on another tweet corpus that had binary sentiment labels, positive, or negative. Then we trained a simple logistic regression classifier on the paragraph-vectors. We found that our classifier had an accuracy of 70% on a validation set.

We then used this classifier to classify our PrEP related tweets into positive or negative labels. We reported the most positive, and most negative tweets, by log(probability), on our full dataset, and on tweets that specifically mention either 'prep' or 'truvada' (Table 4). It seems that overall the classifier seems to capture the sentiment relatively well, however it clearly fails to recognise the sarcasm present in the most positive overall tweet.

The sentiment results aren't perfect, but they do help us identify positive individials in PrEP and HIV public health such as user @greg0wen. The sentiment tweets also let us uncover some of the prevailing stigmas and negative sentiments surrounding PrEP usage.

3. CONCLUSIONS

Conclusions go here.

Example citation (needed right now to compile):[1]

4. ACKNOWLEDGMENTS

Acknowledgements go here.

5. REFERENCES

[1] L. Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

5.1 References

Table 3: Topics Present in 500 HIV/PrEP related Timelines

Topic ID	Word1	Word2	Word3	Word4	Word5
0	0.002*que	0.001*por	0.001*para	0.001*milan	0.001*een
1	0.001*yg	0.001*ini	0.001*yang	0.000*aku	0.000*breakingbad
2	0.001*lmao	0.001*nigeria	0.000*syria	0.000*nigga	0.000*kca
3	0.000*ntv	0.000*inspiringthinkn	0.000*ktnkenya	0.000*haber	0.000*tuscany
4	0.003*gurmeetramrahim	0.001*ang	0.001*ng	0.001*ji	0.001*ako
5	0.001*remedies	0.000*tw	0.000*mx	0.000*rid	0.000*momlife
6	0.000*flipboard	0.000*mongolia	0.000*gettingtozero	0.000*blackburn	0.000*occupy
7	0.002*newsbreakslive	0.002*drudgereport	0.000*woof	0.000*und	0.000*der
8	0.001*giveaway	0.001*xx	0.001*photography	0.001*startups	0.001*anc
9	0.001*realdonaldtrump	0.001*lgbt	0.001*hiring	0.001*uniteblue	0.001*tedcruz

Table 4: Example Positive and Negative Sentiment Tweets

Description	Tweet Body
Most Positive Overall	#apple hey, apple, with your new headphone hack you are in the gouger camp of shkreli who hiked the hiv drug. nice going.
Most Positive PrEP	@greg0wen hi greg, I'm a hiv doctor and health writer doing a piece on hivprep could i poss contact you for some info? thanks, verity x
Most Positive Truvada	great skypeing the john grant just now about bowie , icelandic sagas and truvada! you're right elliot_rose about the song disappointing!
Most Negative Overall	hiv sucks big time!! and I f#n hate it!! We're almost there until this one stupid infection brought him back to where he started. sh#t!!!
Most Negative PrEP	rt @SamNyembe can we attribute hivaids high rate in kzn to many men having not removed their prepuce?
Most Negative Truvada	this #truvada convo has me catching feelings coz majority of men feel 0 responsibility for their reproductive & sexual health