Tools And Techniques For Capturing Possible Hiv Risk-Related Tweets Of Filipinos

Alma Theresa D. Manaloto, Rodolfo C. Raga Jr.

Abstract: Although the number of HIV cases in the Philippines has been steadily increasing, limited studies have been conducted to mitigate the root cause of this problem. The Department of Health expressed the need to design intervention programs to monitor this epidemic. In this study, researchers explored emerging technologies to generate new data that can be used for further studies. A focus group discussion (FGD) among Persons Living with HIV (PLHIV) was conducted to determine how Filipino Men Having Sex with Men (MSM) communicate with each other online. The most common words produced from the FGD were presented to HIV domain experts and validated. In order to collect quantitative data, tweets were extracted to generate profiles of online interactions. Tweepy, a Python library used in accessing the Twitter Advanced Programming Interface (API), was used to collect tweets and the bounding box tool was used to filter tweets coming from the Philippines. The researchers acknowledged that majority of MSM Twitter users preferred to disable geolocation, other techniques were applied to capture risk-related tweets from Filipino MSMs. MySql was considered to handle the tweet repository. A total of 206,822 tweets were extracted from October 8, 2019 to November 6 of the same year. Among all the tweets collected, there are significant amount of tweets that indicate risk related to HIV. Results indicate that Twitter can be utilized to produce data that the government can use to identify high risk locations that required more attention in terms of HIV intervention.

Index Terms: HIV, MSM, Python, Twitter, social media, geolocation, streaming API

1 INTRODUCTION

As the global HIV epidemic continue to pose a significant threat to the general public, it is vital that studies focusing on identifying the causes of the widespread increase of infections are conducted [10][26]. According to the UNAIDS report, the Philippines is the highest in the Asia-Pacific region in terms of increase in the estimated persons living with HIV (PLHIV) from 2010 to 2016 [21]. The main driver of HIV transmission is from Men having Sex with Men (MSM) with a percentage of 84% (44,929) from the total key population [9]. The report states that it was in 2007 when the predominant mode of transmission switched from heterosexual to homosexual - MSM interaction with an alarming rate [12]. In a span of 35 years, 3,076 deaths were reported among people with HIV and 91% of them were males. See Table 1 Demographic date of reported deaths among people with HIV.

The proportion of mortality rate among people with HIV by mode of transmission as seen in Figure 1, signifies that half of the percentage belongs to Male-Male Sex transmission followed by intercourse with both males and males, male to female sex. This study aims to determine the possibility of collecting tweets that can be identified as HIV risk-related communication among key populations of HIV in order to provide useful data to the Department of Health and aid them in planning/designing intervention and monitoring programs.

![Fig. 1. Proportion of Mortality Rate among People with HIV](https://www.doh.gov.ph/")

Intervention programs were conducted to prevent HIV transmission such as: (1) improvement of the knowledge, attitudes, and behaviors through school-based programs [1][24]; (2) 4-hour information dissemination program for sex workers [24][30]; (3) training of HIV social workers [16][24], (4) training of HIV peer leaders of male populations in the Philippines [15][24]. To combat this disease, the Disease Prevention and Control Program (DPCP) under the Department of Health (DOH) already initiated several programs such as:(1) the engagement of 139 Community Health Workers in 39 Global Funds to Fight AIDS, Tuberculosis and Malaria (GFATM)-supported cities; (2)

<table>
<thead>
<tr>
<th>Demographic Data</th>
<th>Jan 2019</th>
<th>Jan 1984 – Jan 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reported Deaths</td>
<td>22⁴</td>
<td>3,076⁴</td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>2,784</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>292</td>
</tr>
<tr>
<td>Age group:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;15 y/o</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>15-24 y/o</td>
<td>4</td>
<td>470⁴</td>
</tr>
<tr>
<td>25-34 y/o</td>
<td>13</td>
<td>1,519⁴</td>
</tr>
<tr>
<td>35-49 y/o</td>
<td>4</td>
<td>890⁴</td>
</tr>
<tr>
<td>50 y/o &amp; above</td>
<td>1</td>
<td>206⁴</td>
</tr>
</tbody>
</table>

¹ Based on the date reported, the actual date of death may not necessarily fall in this reporting month.
² No data available on age for 2 cases.

- Alma Theresa D. Manaloto is a full-time faculty member of Holy Angel University and currently pursuing doctor of information technology degree program in Angeles University Foundation, Angeles City, Philippines. E-mail: almam102@gmail.com
- Rodolfo Raga Jr. is a full-time faculty member of Computer Studies and Engineering department of Jose Rizal University and has PhD in Computer Science from the De La Salle University, Philippines. E-mail: rodolfo.raga@jru.edu
support for the implementation of Community-based HIV Screening (CBS); rapid HIV Diagnostic Algorithm (rHIVDA); (3) establishment of 85 HIV Treatment hubs (including Primary HIV Care Clinics – Social Hygiene Clinics); (4) provision of condoms and HIV test kits; supported the establishment of Community Based Organization (CBO) one-stop shop community centers/sundown clinics; and (5) partnership with other stakeholders on technologies to inform millennials (Safeymph, Battle in the Blood, Safepaces)[11]. DPCP also received support from other government agencies such as: HIV modules in the Comprehensive Sexuality Education (CSE) of DepEd; Policies on HIV in the workplace (DOLE, Civil Service Commission); financial assistance package for PLHIV (DSWD); and Ratification of the Amended HIV Bill in both the House of Representatives and Senate [11]. To monitor new HIV cases and progress of PLHIV, DOH relies on the support of HIV/AIDS Core Teams (HACT). HACT is composed of health workers trained to create, implement, and coordinate policies associated with the diagnosis and care of PLHIV and to aid in the prevention and control of the infection in the hospital. As of this writing, DOH provided certifications to 85 HIV treatment hubs strategically located all over the country to assure that PLHIV can avail services related to HIV treatment and management. DOH Secretary Francisco Duque III believes that a change in mindset must be done to properly fight this disease. The stigma that HIV is equivalent to a death sentence should be corrected and deaths should be averted to treatment [9]. Dr. Gundo Weller, WHO Representative to the Philippines stated that the government along with the WHO and other non-governmental organizations should step up their efforts to reduce the new cases in the country. The government has implemented its new program extending its tailored services to MSM and TGW. It prioritized opening up new clinics in 117 cities that has reported 80% new infections. HIV Clinics are considered a one-stop shop where PLHIV can receive counseling, laboratory work-up, prevention, and treatment services [12]. Beyond these efforts, however, at the current, there is still limited research that has examined the behavior and social activities of this community [6]. Due to high levels of societal stigma, MSM individuals often avoid disclosing their identities thus hindering opportunities to collect data necessary for this type of research [25]. At the same time, traditional approach for collecting data is not only expensive but also time consuming. As such, the lack of data is a major obstacle for developing effective methods for diagnosing, monitoring, and educating potential victims. Online social networking sites such as Twitter can provide an alternative platform for collecting data needed for studying this difficult-to-reach community. A major component of HIV research focused on MSMs requires the study of how an individual acts, how they communicate, what activities they connect and engage in with each other. The rise of Social media use plays a significant role in the development of sexual subculture of MSMs [31]. The most commonly used micro-blogging sites are Facebook and Twitter. It gives freedom of expression by allowing users to share their sentiments, views, activities, etc. The data that is posted on these networks are unstructured data that can be used for qualitative researches specifically big data and can also provide psychological information on a person’s behavior [17][19][34]. A recent study also stated that MSMs, specifically American and Latino men, were comfortable in sharing private information on the internet [33]. This scenario was not just exclusive to a specific geographical area since most of local research validates this claim [1][16].

Twitter is a social media platform where millions of users freely share their sentiments, insights, and activities. Given the amount of information that is shared in this media, specifically filtered data can provide researchers with a real-time source of information that can be analyzed to extract psychological information on attitudes and behaviors not excluding health-related behaviors on any given community [17][19][34]. Thus, Twitter seems like a perfect media for collecting risk-related communications among Filipino MSMs and overall trends in the population [32]. Although there are other social networking applications such as Grindr that is popular to MSMs, the limitation depending on the type of account and the 20-100 mile radius makes Twitter more accessible to the users [14].

Grindr is a dating app popular to the LGBT community but with limited close-proximity search. The terms/keywords that were identified by the respondents as the usual hook-up tagline were classified into themes which include: use of drugs; group sex; sexual preferences; and LGBT slang. Although there are available APIs for Grindr, these are all unofficial and violates the privacy policies of the social media application [23]. This study identified risk-related communications from volunteer PLHIVs and employed emerging data collection tools and techniques on social media, specifically Twitter. Limitations such as tweets without geolocations were also considered to provide a wider sample of risk-related tweets.

2 METHODS

The main goal of this research study was to develop techniques that would employ current tools and technologies in collecting and capturing risk-related communications among social media users in the Philippines specifically MSMs. Figure 7 illustrates the conceptual framework of the study and the activities that were accomplished for the completion of the research work:

2.1 Focus Group Discussion

Focus Group Discussions (FGD) are a form of qualitative research where a group of people sharing common backgrounds or experiences discuss a specific topic, in this case, their behavior in social media [3]. In this study, FGD will be composed of a minimum of 8 and maximum of 15 respondents. A single session with a 30-minute duration is sufficient to answer the five questions prepared for the discussion [22]. Question types for the FGD will be Grand Tour, where the moderator encourages the respondents to speak freely about their experiences; Experience, where it elicits particular event that makes it significant in the topic; and Native-Language, where the researcher aims to learn how the respondents use certain words [8]. FGD is facilitated by a moderator in an unstructured discussion to encourage openness. The researcher chose to conduct focus group discussions in order to determine commonly used words among PLHIV. With the assistance of DOH-Regional Office 3, individuals who are Filipino, 18 years old and above, maintains one or more social media accounts, and used social media for dating or hookups, are recruited to join the discussion. However, users who opted to make their profile private may be excluded from the pool of potential respondents. FGD was held at the Department of Health (DOH) Regional Office in Maimips, City of San Fernando, Pampanga. In order to address juridic/deferential vulnerability, the respondents were assured that they are free to decline the invitation to join the
focus group discussion and will not affect the services he/she receives from the DOH. FGD is documented using a recorder where participants are made aware of. The guide question used in the discussion sought to identify the practices of MSMs online and their preferred social media site in hooking up. The recordings were transcribed and most frequent words that were associated with HIV risk-related communication was validated by HIV domain experts from the DOH. The purpose of validating the keywords from the FGD is to ensure that the keywords that were used to filter tweets were HIV risk-related tweets for accurate tweet classification. In Young's paper on automating HIV identification, he sought the expertise of HIV health workers to code and classify the collected tweets [33]. The analysis of the transcription involves the use of the thematic framework approach where each theme was coded manually [7].

2.2 Twitter Data Collection Tools

In this study, APIs access to Twitter will use the Streaming API method where it provides a continuous stream of public information from Twitter [13]. APIs provides tools for programmers to reuse programs frequently done. It is a collection of libraries, frameworks, and other software development kits to provide low-level access to system resources [18]. To gain access and collect tweet data from Twitter, in this study, an API Streaming method was utilized to provide access to a continuous stream of public information from Twitter [13].

The following are the processes undertaken to collect data from Twitter:

2.2.1 Secure Access to Twitter API

An access token will be supplied upon signing up on the Twitter’s Developer page. Rate limits will be provided depending on the type of account being applied [13]. Twitter has three types of Search APIs: standard, enterprise, and premium search. Standard search API returns a number of tweets based on the provided keyword. This API type is rate-limited and can only handle 180 user authentication requests per 15 minutes. On the other hand, Enterprise and Premium search API has two types: 30-Day Search API and Full-Archive Search API. The 30-day search API can only produce tweets within the last thirty days and the latter produces tweets posted as early as 2006. Enterprise and Premium has the same types, methods and parameter but differs in terms of latency and full-fidelity. Premium search APIs provides better query results from the Tweet archive due to low-latency and full fidelity. Enterprise and Premium has subscription fees while Standard is free of charge [5]. For this study, only Standard access was used for testing purposes.

2.2.2 Install XAMPP Apache Server

XAMPP is an open-source web-server solution that consists of Apache HTTP Server, MariaDB, and interpreters for PHP and Perl programming languages. To install a Windows’ version of XAMPP, the local machine should have Microsoft Visual C++ Redistributable[20][28].

2.2.3 Setup localhost database for storage of collected tweets

A local database was provisioned to store collected tweets. The structure of the database includes the following metadata:
2.2.5 Install Tweepy
Tweepy is an open-source Python package that allows developers to access Twitter’s API using Python. It handles implementation details of API endpoints such as data encoding and decoding, HTTP requests, results pagination, OAuth authentication, rate limits and Streams. To install Tweepy, use Python’s package manager pip [29].

2.2.6 Create Python code and Consume API
The code will start with importing the necessary methods from the tweepy library such as StreamListener, OAuthHandler, Stream, json, and requests.

```python
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import json
import requests
```

Fig. 4 Screenshot of Code Snippets

After importing the necessary libraries, assign values to the variables that hold the user credentials. Next step is to add a listener that posts received tweets to the local database. Final part of the code involves user authentication before connecting to Twitter and applying the filter criteria in this case, the coordinates of the bounding box.

```python
l = StdOutListener()
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
stream = Stream(auth, l)
stream.filter(locations=[116.72, 5.09, 129.28, 19.83])
```

Fig. 5 Screenshot of Filter by Location

2.2.7 Run the Python Code and Call API
Before running the code, make sure that the localhost is running by checking the XAMPP control panel. Upon compiling the code, it will automatically pass and store the tweets on the database along with the backup text file.

2.2.8 Twitter Data Collection with Filter Keywords
Two types of python codes were developed to accommodate filter criteria of location and keywords. It was found that the mentioned criteria cannot be applied simultaneously so a separate run for each code were implemented. The following figure shows the screenshot for the second file with the keyword as filter criterion:

```python
l = StdOutListener()
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
stream = Stream(auth, l)
stream.filter(track=['alterparty', 'alterph', 'alterbagets', 'alt`
```

Fig. 6. Screenshot of Filter by Keyword

2.3 Ethical Considerations
All information gathered from the data gathering activities such as interviews, focus group discussions, and tweets were kept confidential and the identity of the individuals was treated anonymously. An informed consent and non-disclosure agreements were accomplished by the researchers and the participants of the study. An FGD guide question made sure that the focus group discussion was properly implemented and sensitive to the respondents' feelings.

3 RESULTS
The focus group discussion with twelve PLHIVs described how Filipino MSMs use different Social Media Sites and their common practices. All of the twelve respondents maintained at least three social media accounts and updates/posts at least once a day and at most ten to fifteen times every day. Sixty six percent of the respondents stated that although they frequently posts updates on their Facebook accounts, they are most likely to posts their true sentiments and personalities on Twitter. According to them, Twitter has a lenient policy in terms of content and it is in Twitter where they can create alter accounts. Alter accounts or pseudonym are created to hide the identity of individuals. True identity can be disclosed through direct messaging, or if the user would decide to do so. Fifty eight percent of the respondents also disclosed that they used Twitter to hook up with someone despite having other apps such as Grindr. Figure 8 shows the keywords identified by the respondents and number of instances the word was seen on tweets:

Fig. 8 Keyword Instances in Tweet Dataset

From the two separate data collection process: filtered by location and filtered by keyword, a total of 206, 822 tweets...
were captured. Tweets collected via location filtering reached 136, 276 in a span of 30 days and a total of 70, 546 for tweets with keyword filtering. The geolocated tweets was able to capture 404/136,276 risk-related tweets while the latter has 4,075/70,546 risk-related tweets.

4 CONCLUSION

Based on the dataset consisting of tweets collected using two methods, it is important to note that there is a higher probability that promiscuous social media users, having 5.58% of risk-related tweets with disabled location compared to 0.30% risk-related tweets with geolocation, prefer to disable their geolocation settings which may hinder identifying their location. Results show that there’s a higher viability of collecting risk-related tweets using filtered keywords rather than according to location. Since the aim of this study is to provide the DOH information on where they should focus their intervention programs, this problem should be addressed. It is suggested that alter hashtags be used to identify user’s location in cases where geolocation was disabled. Alter hashtags in Twitter are posted along with the user’s location, e.g. #altercebu, #altermakati, and #alterbulacan. This study was able to establish that a wider initiative on developing programs for HIV intervention should be done in order to support the programs of the government. Since there is a data available on these platforms, deep exploration on big data should be developed to predict future HIV infection. Some keywords identified by respondents can only be classified as risk-related communication depending on the context of the tweet. Therefore, manually classifying those tweets is time-consuming. It is recommended that to capture all tweets coming from the Philippines, developer’s account should be upgraded to Premium and machine learning be applied in this study to rapidly identify/classify risk-related tweets. This will enable wider coverage of data that can be processed for deeper analysis.

5 ACKNOWLEDGMENT

The authors wish to thank the support of the Department of Health Regional Office 3, specifically the Communicable Disease Cluster, for the assistance in facilitating the focus group discussion and allowing the authors to conduct interviews among social health workers.

6 REFERENCES


[3]. Colucci, E., 2007. "Focus groups can be fun": The use of activity-oriented questions in focus group discussions. Qualitative health research, 17(10 pp.1422-1433


[7]. Gohil, J., Are the Philippines ready for HIV self-testing? (Doctoral dissertation, Department of International Public Health, Liverpool School of Tropical Medicine).


[21]. PHILIPPINES ADDRESSES RISING TREND IN NEW HIV INFECTIONS: Department of Health website. (n.d.).


