Title
PIRCNET: A Data Driven Approach to HIV Risk Analysis

Permalink
https://escholarship.org/uc/item/8q79x6pr

Author
Desai, Purvi Jayesh

Publication Date
2016

Peer reviewed|Thesis/dissertation
PIRCNET: A Data Driven Approach to HIV Risk Analysis

A Thesis submitted in partial satisfaction of the requirements for the degree of Master of Science

in

Computer Science

by

Purvi Jayesh Desai

Committee in charge:

Nadir Weibel, Chair
Amarnath Gupta
Lawrence K. Saul

2016
The Thesis of Purvi Jayesh Desai is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego
2016
DEDICATION

To Mummy, Pappa, Ba, and Dada

It’s Done!
Every contact leaves a trace. This is evidence that does not forget.
   It is not confused by the excitement of the moment.
   It is not absent because human witnesses are.
   It is factual evidence.
Physical evidence cannot be wrong, it cannot perjure itself,
   it cannot be wholly absent.
Only human failure to find it, study and understand it,
   can diminish its value.

*Dr. Edmond Locard’s Exchange Principle*
*Contributed by Jayesh Desai*
TABLE OF CONTENTS

Signature Page ................................................................. iii
Dedication ................................................................ iv
Epigraph ..................................................................... v
Table of Contents ................................................................. vi
List of Figures ................................................................. viii
List of Tables ................................................................ x
Acknowledgements ............................................................... xi
Abstract of the Thesis .......................................................... xii

Chapter 1  Background and Motivation ................................. 1
  1.1 HIV: Overview ............................................................. 1
    1.1.1 HIV Transmission .................................................... 2
    1.1.2 How HIV affects different groups of people ............... 2
  1.2 Intervention and Prevention, and Challenges ................. 3
    1.2.1 Challenges ............................................................. 4
    1.2.2 Augmenting Intervention with Social Network Information . 5
  1.3 Research Contributions at UC San Diego: PIRCNET ......... 5

Chapter 2  PIRCNET ............................................................. 7
  2.1 The Premise ................................................................. 7
  2.2 Data Collection ............................................................ 8
  2.3 Data Cleaning .............................................................. 9
  2.4 PIRCNET’s Data Model ................................................. 11
  2.5 Acknowledgements ...................................................... 14

Chapter 3  Introducing Machine Learning Techniques to Better Classify at-risk Users ............................................... 15
  3.1 Data and Observations .................................................. 16
    3.1.1 Data Labeling ......................................................... 16
    3.1.2 Labeling Data with Domain Experts ......................... 16
    3.1.3 Initial Observations ............................................... 18
  3.2 Text Based Classification of Tweets ................................. 19
  3.3 Experiments and Results ................................................. 22
    3.3.1 Naive Bayes Classification ....................................... 22
    3.3.2 Logistic Regression Based Classification .................... 25
3.3.3 Support Vector Machine (SVM) Based Classification .......................... 27
3.4 SVM: Tuning the Hyperparameter ............................................... 31
  3.4.1 ROC Curve Overview ......................................................... 31
3.5 Classifying Users for HIV risk based on their Tweets’ Risk .................. 33
  3.5.1 User Risk Analysis with our Labeled Data ............................ 34
  3.5.2 Risk Prediction with the Global Corpus ............................... 34

Chapter 4 Exploring the HIV at-risk Network ...................................... 38
  4.1 User-User Connections Derived from the HIV Risk Network .......... 38
    4.1.1 Conversations .............................................................. 40
    4.1.2 Geographic + Temporal Co-location ............................... 46
    4.1.3 Mentions ................................................................. 49
    4.1.4 Follows ................................................................. 52

Chapter 5 Augmenting Text-based HIV Risk Classification with Network Information .......................................................... 58
  5.1 Premise ................................................................. 58
  5.2 Proposed Model .......................................................... 59
  5.3 Experiments ........................................................................ 63
    5.3.1 Conversations .............................................................. 63
    5.3.2 Co-location ............................................................... 65
    5.3.3 Mentions ................................................................. 68
    5.3.4 Follows ................................................................. 70
  5.4 User Risk Score Analysis .............................................................. 73
    5.4.1 Comparing Network Only with Text Only Analysis ........... 74

Chapter 6 Inferences and Directions for Future Work ......................... 76
  6.1 Summary of Inferences ....................................................... 76
  6.2 Challenges ................................................................. 78
  6.3 The Road Ahead ............................................................. 80

Appendix A HIV Risk Words .............................................................. 82

Bibliography .................................................................................. 83
LIST OF FIGURES

Figure 1.1. “HIV Diagnosis by Transmission Category” 3
Figure 2.1. “Data Collection Architecture” 10
Figure 2.2. “Data Model of the HIV at-risk social network” 13
Figure 3.1. “Data Labeling Interface” 17
Figure 3.2. “Tweet Score Frequency Distribution” 18
Figure 3.3. “Cumulative Tweet Score Frequency Distribution” 19
Figure 3.4. “Logistic Function” 26
Figure 3.5. “ROC Curve for Text Based HIV Risk Classification with SVM” 32
Figure 3.6. “HIV risk visualization for users based on the risk type over one month” 36
Figure 4.1. “Graph Visualization of Conversations within the HIV Risk Network” 42
Figure 4.2. “Graph Visualization of Homosexual Conversations within the HIV Risk Network” 43
Figure 4.3. “Zoomed View: Graph Visualization of Conversations within the HIV Risk Network” 44
Figure 4.4. “Frequency Distribution of Node Degrees within the Conversations Network” 46
Figure 4.5. “Graph Visualization of User-User Connections based on Co-location” 48
Figure 4.6. “Zoomed View: Graph Visualization of User-User Connections based on Co-location” 49
Figure 4.7. “Frequency Distribution of Node Degrees within the Co-location Network” 50
Figure 4.8. “Graph Visualization of User-User Connections based on User Mentions in Tweets” 52
Figure 4.9. “Zoomed View: Graph Visualization of User-User Connections based on User Mentions in Tweets” ............................. 53
Figure 4.10. “Frequency Distribution of Node Degrees within the Mentions Network” ............................................................... 53
Figure 4.11. “Graph Visualization of User-User Connections based on Follows” ................................................................. 55
Figure 4.12. “Zoomed View: Graph Visualization of User-User Connections based on Follows” ......................................................... 56
Figure 4.13. “Frequency Distribution of Node Degrees within the Follows Network” ................................................................. 57
Figure 5.1. “Social Network Structure with Tweets and Users” ............. 60
Figure 5.2. “Classification Accuracy with Network Features: Direct Conversations” ................................................................. 65
Figure 5.3. “Classification Accuracy with Network Features: Indirect Conversations” ................................................................. 66
Figure 5.4. “Classification Accuracy with Network Features: Direct Colocation” ................................................................. 67
Figure 5.5. “Classification Accuracy with Network Features: Direct Colocation and Direct Conversation” ................................................. 68
Figure 5.6. “Classification Accuracy with Network Features: Direct Mentions” ................................................................. 70
Figure 5.7. “Classification Accuracy with Network Features: Direct Conversations, Colocation, and Mentions” ................................................. 71
Figure 5.8. “Classification Accuracy with Network Features: Direct Follows” ................................................................. 72
Figure 5.9. “Classification Accuracy with Network Features: Direct Conversations, Colocation, Mentions, and Follows” ................................................. 73
Figure 5.10. “Classification Accuracy with Network Features: Random Connections” ................................................................. 74
Figure 5.11. “User Risk Score: Network vs Text” ................................................. 75
LIST OF TABLES

Table 3.1. Naive Bayes Classification Accuracy for HIV at-risk Tweets . . . . . . . . . . . . 25
Table 3.2. Logistic Regression Classification Accuracy for HIV at-risk Tweets 27
Table 3.3. SVM Classification Accuracy for HIV at-risk Tweets . . . . . . . . . . . . 30
Table A.1. HIV Risk Words . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
ACKNOWLEDGEMENTS

It has been an incredible journey with so many people I am grateful to. To begin with, I would like to express my sincere thanks to Nadir Weibel for being an immense source of inspiration and encouragement all along. His outlook towards research and passion for this project was what kept me going. I would also like to thank Amarnath Gupta and Lawrence Saul for guiding me at every step with their profound knowledge and expertise. I could not have imagined having a better team for my research work.

I would like to thank Narendran Thangarajan for being an excellent mentor. Working with him never ceased to be exciting. I would also like to thank Dr. Susan Little and Dr. Nella Green for their guidance and insightful comments.

I cannot even begin to properly thank my family whose support and advice has been instrumental to everything I have done and will do. My friends, who have always been there to help and encourage me. A special thank you to Jayesh Desai and Gautam Akiwate for always being there to help, encourage and inspire, and for patiently helping me revise my thesis. I couldn’t have done this without your support.
The world continues to battle with HIV, one of the major public health issues. It has no known cure yet and has claimed several million lives so far. Over a period of time, there has been tremendous progress in intervention methods to prevent the spread of this disease. Traditional intervention schemes have however been faced with several challenges. An important challenge is the lack of a real-time radar to address the most at-risk communities at a given time and location. Another challenge is related to limited HIV prevention resources. Such challenges makes it important to progress towards a more targeted, evidence based, and data-driven intervention method. Along with being
data-driven, such intervention also needs to be effective, scalable, and tailored as per the distinct needs of communities.

The primary goal of this thesis is to unveil HIV transmission networks with the help of information gathered from digital social media footprint. Recent research in this area suggests the feasibility of using Twitter as a platform to uncover HIV at-risk behaviors among communities. Our research further explores this by trying to understand how likely a person may be at the risk of acquiring or transmitting HIV based on the nature of the content shared or consumed via Twitter. We focus on the quality of the data gathered, as well as its temporal and spatial dimensions to help with a more real-time risk analysis.

To begin with, we look only at the text-based information shared by users to design a model for HIV risk prediction. We move on to improving the risk analysis model by exploiting social relationships from network induced user-user connections in addition to the text-based information. Towards the end we discuss how our model can be used to inform the current intervention methods, followed by areas for improvement and future directions to further refine our techniques.
Chapter 1

Background and Motivation

In this chapter, we discuss at a high level, the premise of our research towards a data-driven approach for HIV intervention and prevention schemes. It begins by outlining the basics of HIV, its routes of transmission, and how it affects different demographics. It then goes on to discuss the importance of targeted HIV intervention, its challenges, and our contribution to the research that aims at tackling these challenges with a data-driven approach.

Recent research shows that HIV continues to spread despite its restricted routes of transmission. Our goal is to tap into the reasons for HIV spread that may go unchecked and proceed towards targeted, evidence-based intervention.

1.1 HIV: Overview

HIV stands for *Human Immunodeficiency Virus*. As the name suggests, HIV weakens a person’s immune system by attacking specific immune cells which are responsible for protecting the person from infection. Weakening of the immune system makes a person susceptible to infections from other viruses and bacteria.

No effective cure exists for HIV. It can be transmitted via the exchange of a variety of body fluids from infected individuals, such as blood, semen, and vaginal secretions. However, the spread of HIV can be curbed with appropriate control measures. That said,
making these measures accessible to the at-risk populations still remains a big challenge. Given how HIV spreads, some groups of people are more likely to acquire HIV than others as a result of many factors including their sex partners and their own risk behaviors [3][10].

HIV continues to be a major global public health issue, having claimed more than 34 million lives so far. While the UNICEF reports dwindling numbers of HIV infections, it states that “we must do more to break the deadly cycle of transmission”. According to the 2015 Statistical Update on HIV and AIDS among Children, HIV/AIDS is the second leading cause of death for teens globally, and AIDS-related deaths among adolescents have tripled since 2000 [5].

1.1.1 HIV Transmission

As mentioned above, HIV can be transmitted via the exchange of a variety of body fluids from infected individuals. Primary methods of transmission are [3]:

1. Having anal or vaginal sex with someone who has HIV without using a condom or taking medicines to prevent or treat HIV. Anal sex is the highest-risk sexual behavior. Vaginal sex is the second-highest-risk sexual behavior.

2. Sharing needles or syringes, rinse water, or other equipment used to prepare drugs for injection with someone who has HIV.

1.1.2 How HIV affects different groups of people

If we look at HIV diagnoses by how people acquired the virus (transmission category) in Fig. 1.1, we see that gay and bisexual men are most at risk. In 2014, gay and bisexual men accounted for 67% while individuals infected through heterosexual sex accounted for 24% of all new HIV diagnoses [3].
Figure 1.1. HIV Diagnosis by Transmission Category

The spread of HIV was earlier restricted only to the high risk populations like sex workers, men having sex with men (MSM), and Intravenous Drug Users (IDUs) due to the restricted routes of HIV transmission. However, it started spreading to the general population through a network that comprised of clients of sex-workers and bisexual men. This led to an overall increase in the spread of HIV in the general population.

1.2 Intervention and Prevention, and Challenges

The good news is that over a period of time there has been considerable progress in bringing these numbers down. According to the Global Health Observatory (GHO) data: “Expanded access to anti-retroviral therapy (ART) and a declining incidence of
HIV infection have led to a steep fall globally in the number of adults and children dying from HIV-related causes. The estimated 1.5 million people dying from HIV globally in 2013 were 22% fewer than in 2009 and 35% fewer than when the number peaked in 2005. Children (younger than 15 years) in 2013 had 31% fewer deaths from HIV compared with 2009 and 40% fewer deaths compared with 2005.”

The most common methods of direct intervention for HIV prevention today are administering Pre-Exposure Prophylaxis (PrEP), distributing free condoms, and new needles/syringes. Other means of intervention include HIV transmission and prevention awareness campaigns. However, in order to continue to keep a check on the spread of HIV, it is important to continue to strive towards a more targeted, evidence-based HIV intervention plan; not only to intensify the prevention efforts, but to also reduce HIV related stigma that acts as a major barrier for HIV infected patients to seek help from HIV prevention and treatment programs.

1.2.1 Challenges

HIV intervention schemes face several challenges. To begin with, only few people with HIV are actually aware of their infection. Many infected people do not proactively seek or receive any form of treatment. This also has to do with the nature of this disease. HIV doesn’t show up in the test results right away. It can take up to 3 months after being exposed to HIV for infected individuals to test positive. To add to this, diverse populations require tailored prevention approaches with social and economic disparities leading to a disparity in the HIV rates. The limited amount of prevention resources forces difficult choices where people at the highest risk of HIV may not receive help in time. This is one of the important reasons why HIV intervention strategies need to be more targeted and evidence-based.
1.2.2 Augmenting Intervention with Social Network Information

The advent of social networks has made it feasible to characterize communities owing to their real-time communication and information dissemination capabilities. Recent research has outlined the feasibility of utilizing social networks for surveillance of HIV infections [13]. In our effort to have more efficient public health tools capable of targeted intervention and prevention, the big question is if digital social networks can be exploited effectively to identify populations vulnerable to HIV.

Related Work: A team of 3 researchers headed by Dr. Sean D. Young at UCLA conducted a study [16] on Twitter data to show that social network information could help perform epidemiological analysis. This could further help identify early warning signals for HIV risk. Another study was conducted by researchers at Microsoft Research, Redmond about people who are trying to quit smoking. This study further helped corroborate that people’s behavior can be correlated to their social media footprint.

1.3 Research Contributions at UC San Diego: PIRC-NET

As discussed in the sections above, it is important to investigate more effective means of targeted, evidence-based HIV intervention strategies. As a step towards this goal, former Graduate Student Researcher at UC San Diego, Narendran Thangarajan under the guidance of Dr. Nadir Weibel, Dr. Amarnath Gupta, Dr. Susan Little, and Dr. Nella Green worked towards a new method of characterizing and identifying HIV at-risk populations. This effort focused on the local population of San Diego County and used publicly available social media data as an indicator for HIV risk. The idea was to apply social network analysis to combine real-time Twitter information extracted from the San Diego area with the data available as part of the Primary Infection Research Consortium
(PIRC) at the AntiViral Research Center (AVRC) at UCSD and evaluate opportunities to target HIV testing and prevention efforts to the communities at the greatest potential risk of HIV acquisition.

The next chapter discusses research that is done as part of the preliminary analysis with PIRCNET in more detail. The goal of this work was to understand how we could better characterize and identify vulnerable populations using tweets, social relationships, and related information inferred from Twitter’s social graph.

PIRCNET provides the base infrastructure that gathers and filters tweets that are potential candidates for HIV risk analysis. It also gathers the corresponding Twitter users and relevant network based data from Twitter’s social graph such as follower/followee connections, retweets, and mentions. With the inherently noisy and sparse nature of the gathered tweets that reduces the signal to noise ratio in the collected data, it was important to devise methods to improve the signal in the data. This thesis builds on top of the previous research done as part of PIRCNET and is primarily focused towards improving the signal in the data gathered. Specifically, it focuses on the application of machine learning techniques to refine the categorization of HIV risk tweets and corresponding users by analyzing the textual content. It also explores how the augmentation of this textual content categorization with network information might help further refine the categorization.
Chapter 2

PIRCNET

2.1 The Premise

We discussed in Chapter 1 how HIV’s primary route of transmission is governed by connections within social or sexual networks. Analyzing such networks can inform public health measures to contain the spread of HIV. However, these networks tend to have a dynamic structure. Further, HIV has a long incubation period between transmission and disease state. It also has a low transmission rate per contact. Hence, traditional methods of defining network features through interviews and partner tracing would not be as effective for an infection of this nature [9]. Research under PIRCNET aims at exploiting data from social networks that help identify relevant network features and thereby enable us to better characterize transmission networks. Given how people’s social media footprint mirrors their real life to a large extent, real-time analysis of social networks could help build an infection surveillance radar of HIV transmission risk behavior. Twitter data has the potential to equip us with location and population based HIV risk behavior indicators to infer transmission networks which could help provide an early warning indicator for HIV risk [13].

This study was based off of the Twitter networks to understand how online HIV at-risk behaviors relate with real-life behavior. The area of focus was San Diego County.
An infrastructure was built to collect, clean, filter, organize, and manage the stream of
tweets in real-time [15]. This data was cleaned to make it amenable for initial exploratory
analysis to unravel the relationships between different HIV risk behaviors. Various
dimensions of this data such as geography, demographics, and social groups were used
to model HIV risk. This chapter describes at a high level, the infrastructure for data
collection and filtering, data cleaning, and the data model. It also describes how this
infrastructure enables extraction of important information about HIV risk behavior in
San Diego.

2.2 Data Collection

Twitter provides an Application Programming Interface (API) to programmati-
cally access their data. Along with the standard APIs, Twitter provides a Streaming API
that creates a long-standing connection between the client and the server, and streams the
incoming tweets to the subscribing clients.

In order to capture tweets from San Diego alone, Twitter’s *Filter Hose* API was
used. This API allows collecting geo-tagged tweets that are generated within a geocoded
bounding box in real-time.

The tweets collected using the Streaming API are continuously pushed onto a
MongoDB database that stores them and enables easy access for analysis. MongoDB was
used mainly owing to the ease of storing semi-structured documents, high write through-
put, and support for native map-reduce queries for performing on-demand aggregations.

To identify HIV at-risk tweets, all tweets from San Diego County were filtered
to create a smaller corpus of tweets based on the presence of certain HIV transmission
risk words in the tweet content. A “risk word” is essentially a term that is considered
to be positively correlated with HIV risk behavior. Based on their domain expertise,
the clinical collaborators on our research team helped define five broad categories of
HIV risk words. These buckets were populated with words frequently used in the local community of San Diego. A list of these words is listed in Appendix A.

The five broad categories of HIV risk words were as follows:

1. **Drug Bucket**
   
meth, ice, snow, cocaine, party&play

2. **Sex Bucket**
   
creampie, cottaging, bronco, party&play

3. **Sex Venues Bucket**
   
loft, redwing, bourbon street

4. **Homosexual Terms Bucket**
   
homo, gay, queen

5. **Sexually Transmitted Infection Bucket**
   
syfy, drip, gleet

Data collection is followed by data filtering to weed out false positives, and then pulling from Twitter all the related users who either re-tweeted an HIV risk tweet or were mentioned in one. All this data is stored on a separate MongoDB collection. Fig. 2.1 illustrates this process.

### 2.3 Data Cleaning

The data collection process filters the data based on the presence of keywords from the five HIV risk categories. However, on a microblogging social network like Twitter, the text is limited to 140 characters. This leads to the use of a lot of text shortening, abbreviation, and emoticons. Also, capturing the context along with the keywords was important in order to mitigate the overall noise in the data collected.
Figure 2.1. Data Collection Architecture

The data collection process filtered the incoming tweets based on the presence of certain risk keywords. However, these risk words could mean something entirely different in different contexts, and hence at times be misfiltered as at-risk. In order to weed out such false positives, there was a notion of having inclusion and exclusion lists based on the co-occurrence of words with these risk keywords. For instance, the exclusion list for the keyword “crack” (which is slang for meth/drug) would include crack me up, crack myself up, crack up, crack open, crack of dawn. A tweet with a phrase that reads “crack me up” would pass our first crude filter for at-risk tweets based merely on the presence of certain predefined keywords. However, the second level filter based on inclusion and exclusion lists would weed out such false positives based on the co-occurrence of words. The inclusion list on the other hand would include words that if co-occurred with the keywords under consideration would allow the tweets to pass through this second level of filtering.

This data cleaning phase led to the reduction of tweets that passed the initial
keyword filter based by 60%, thus improving the signal in the data. This process of filtering incoming tweets happens as a batch process every three hours.

2.4 PIRCNET’s Data Model

Capturing user connections within the social network alongside analyzing the raw text of the tweets gathered, could provide meaningful signal for HIV risk analysis. The user-tweet and user-user connections were modeled in a graph in the form of nodes and relationships with the help of Twitter’s APIs to access publicly available data. Fig. 2.2 shows an instance of this data model.

The nodes were divided into seven different categories:

1. **USER** nodes correspond to each user in the dataset. Each USER node in our graph is at HIV risk based on their tweets.

2. **TWEET** nodes corresponding to each HIV risk tweet in Twitter.

3. **HASHTAG** nodes correspond to each hashtag used in the HIV risk tweets.

4. **URL** nodes correspond to each URL being referred to in the HIV risk tweets.

5. **FOLLOWER_USER** nodes correspond to the set of users that follow each of the USER nodes. So FOLLOWER_USER may or may not be an HIV risk Twitter user.

6. **ONTOLOGY_BUCKET** nodes correspond to each of the five risk buckets we defined above.

7. **ONTOLOGY_INSTANCE** nodes correspond to each HIV risk word in each HIV risk bucket.

   Each of these nodes were connected via edges that were categorized into nine different types:
1. **TWEETED** edges from a USER node to a TWEET node, indicate the author relationship of the tweet.

2. **IS_REPLY_FOR** edges from a TWEET node to another TWEET node indicate that a tweet is a reply to another one.

3. **RETWEET_OF** edges from a TWEET node to another TWEET node indicate that the tweet is a retweet of another tweet.

4. **FOLLOW** edges from a USER node to a USER node indicate what other users the current user is following on Twitter.

5. **MENTIONED_IN** edges from a USER node to a TWEET node indicate when a user is mentioned (with reference to @ handle) in another tweet.

6. **HAS_HASHTAG** edges from a TWEET node to a HASHTAG node indicate that the specific tweet contains the listed hashtag (#).

7. **HAS_URL** edges from a TWEET node to a URL node indicate if and what URL is included in a Tweet.

8. **HAS_RISK_WORD** edges from a TWEET node to an ONTOLOGY_INSTANCE node indicate the risk word the specific Tweet has been assigned to (can be multiple).

9. **INSTANCE_OF** edges from an ONTOLOGY_INSTANCE node to an ONTOLOGY_BUCKET node indicate the bucket every risk word belongs to.

In order to facilitate the analysis of this graph based data at a network level, this data was stored in a Neo4j graph database\(^1\). The topological information along with the

\(^1\)http://neo4j.com – A highly scalable, native graph database
text based information could potentially help analyze HIV risk from a new dimension of user-user connections. For instance, one could identify “hubs” or “influencers” from the network based on the several different types of social relationships described above, or similarity in HIV riskiness between two users that share a connection. A graphical representation of the risk-network in terms of social network relationships could therefore offer significant added value.

In Chapter 3, we describe in detail the work done to improve the data filtering process by using a text-based supervised learning approach. In Chapter 4, we describe the user-user connections that can be inferred from the umbrella risk network built in this chapter. Chapter 5 further describes how these network connections can be used to augment the text-based classification approach for risk classification to improve the risk

Figure 2.2. Data Model of the HIV at-risk social network
analysis.

2.5 Acknowledgements

This chapter gives an overview of the work done by former Graduate Student Researcher, Narendran Thangarajan, under the guidance of Prof. Nadir Weibel, Prof. Amarnath Gupta, Dr. Susan Little, and Dr. Nella Green. All references in this chapter are taken from the paper titled “Analyzing Social Media to Characterize Local HIV At-risk Populations” [15].
Chapter 3

Introducing Machine Learning Techniques to Better Classify at-risk Users

The two-stage filtering process described in Chapter 2 significantly brings down the number of relevant tweets that are considered at risk. However, with this heuristic approach a fairly large number of false positives still remain in the corpus. Given the large size of the corpus, using Machine Learning techniques is a logical extension to build a better filter as well as to get additional information from the tweets.

As a first pilot to understand the veracity of the filtering process, we had a domain expert on our team assign “Positive” and “Negative” labels to 618 filtered tweets. “Positive” implies that the tweet content exhibits behavior or tendencies that are highly correlated with HIV risk. “Negative” implies that the tweet content indicates very low/no HIV at-risk behavior. From this pilot, we observed that the filtered data comprised of a considerably large number of false positives.

In this chapter, we discuss how the tweets filtered as candidates for HIV at-risk tweets are further classified using standard text-based supervised machine learning techniques to achieve a stronger signal by having a higher number of true positives and a lower number of false positives. In other words, we would like to improve the ratio of true positives to false positives.
3.1 Data and Observations

3.1.1 Data Labeling

In order to obtain a labeled data-set, we initially planned to crowdsourcethe task on a mechanical turk platform where turkers would manually assign labels to the tweets based on their perception of the tweet’s HIV risk. However, we soon realized that lack of sufficient domain knowledge and familiarity with Twitter vocabulary used in the specific context of our Twitter at-risk users would potentially lead to poor labels. To get more reliable labels, we decided to get the data hand-labeled by domain experts only.

In our first pilot, referenced above, of the 618 tweets 171 were marked “Positive’ and 447 as “Negative” for HIV risk behaviors. From this pilot, we observed that the filtered data comprised of 27.6% true positives and 72.4% of false positives. After reviewing the labels and feedback from the domain expert who was labeling the data, we realized a few caveats in the labeling process that led to noisy labels. The first was that a lot of tweets were in Spanish, or had terms that were not decipherable. Restricting the user to label a tweet as “Positive” or “Negative” forced her to pick one of the two, whereas in fact, it would have been useful to have a separate “Do Not Understand” category. This would lead to lesser noise in our labels of interest (“Positive” and “Negative”). We also observed that the tweet content suggested different degrees of risk. Hence, it would be worthwhile replacing binary labels with a multiple-point Likert scale that captures the degree of risk as well. Finally, we also very clearly realized how this method was not scalable with just one person labeling the tweets.

3.1.2 Labeling Data with Domain Experts

To improve our initial data labeling process, we created a data labeling platform with a web interface (shown in Fig. 3.1) wherein a user was displayed the tweet text and
could assign one of the following labels with respect to HIV risk behavior:

- 4: Very High
- 3: High
- 2: Low
- 1: Very Low
- Don’t Understand

A labeling event was organized and 30 participants from the MSM community were recruited to participate in this event. At the labeling event, each participant had to label 200 tweets on the Likert scale described above. These tweets were randomly assigned to the participants. In addition, each tweet was assigned to 3 participants to smooth out any labeling discrepancies. At the end of the event we collected 6000 labeled tweets overall, and 2000 uniquely labeled tweets among them.

Figure 3.1. Data Labeling Interface
3.1.3 Initial Observations

In order to better understand the data we collected, we analyzed the labeled tweets. Fig. 3.2 shows the frequency distribution of the tweets’ scores based on the individual score each tweet got after receiving at most 3 labels from 3 different participants. Most of the tweets have a score of 1 (low risk). This is consistent with the fact that the dataset had a lot of false positives in the first place.

Fig. 3.3 shows the frequency distribution of the tweets based on the cumulative score each tweet got after receiving at most 3 labels from 3 different participants. Most of the tweets have a cumulative score from 3-6. This, again, is confirming the fact that the data set had a lot of false positives.

![Frequency vs Tweet Score (Label) Plot](image)

**Figure 3.2.** Tweet Score Frequency Distribution
The next step in our process was to train a classifier with this labeled data based on a supervised learning approach. The goal was not only to achieve a higher classification accuracy, but also a higher true positives to false positives ratio.

### 3.2 Text Based Classification of Tweets

Our first approach was based on text-only classification of the collected tweets. The goal of the text-based classification of tweets was to categorize them into two distinct classes, i.e. “Positive” HIV risk and “Negative” HIV risk behavior. We treat this as a document classification problem where each tweet represents a document and belongs to a distinct class. With text-based categorization, the objective is to use machine learning to train classifiers from examples so that they would perform the class assignment.
automatically. The first step in text classification is to transform documents, which in our case are strings of characters, into a representation suitable for the learning algorithm and the classification task [6].

For the purpose of our experiments we modeled the tweets features in the following different ways:

1. **Feature Set #1**: Bag-of-words Model (Unigram Model): Each distinct word $w_i$ in a document corresponds to a feature, with the number of times word $w_i$ occurs in the document as its value. The feature vector length for this set of features was 5964.

   Example: Unigrams for “I do meth cz I love meth” are \{I:2, do:1, meth:2, cz:1, love:1\}

2. **Feature Set #2**: Unigrams + Bigrams: Along with unigrams, each pair of words $w_i$ followed by $w_{i+1}$ corresponds to a feature, with the number of times such word pairs occur in the document as its value. This captures the co-occurrence of words with the risk keywords of interest, which is important to achieve fewer false positives. The feature vector length for this set of features was 20,087.

   Example: Bigrams for “I do meth cz I love meth” are \{I do:1, do meth:1, meth cz:1, cz I:1, I love:1, love meth:1\}

3. **Feature Set #3**: Unigrams + Bigrams + Trigrams: Along with unigrams and bigrams, each triplet of words $w_i$ followed by $w_{i+1}$ followed by $w_{i+2}$ corresponds to a feature, with the number of times such triplets occur in the document as its value. The feature vector length for this set of features was 35,910.

   Example: Trigrams for “I do meth cz I love meth” are \{I do meth:1, do meth cz:1, meth cz I:1, cz I love:1, I love meth:1\}
4. **Feature Set #4**: Unigrams + Bigrams + Trigrams + Risk Words: Along with unigrams, bigrams, and trigrams, every term that is considered a risk word corresponds to a feature with the number of times the risk word appears in the document as its value. The feature vector length for this set of features was 36,500.

Example: Risk words are \{crack, gay, homo, meth\}

Our goal was to categorize tweets as “Positive” or “Negative” HIV risk tweets based on the text content using a supervised learning algorithm. For the purpose of learning a classifier we trained our model on the labeled tweets we had gathered from the experts labeling event described above. We segmented our data into two parts, the “Training Set” and the “Testing Set”.

**Training set**: A set of examples used for learning, that is to fit the parameters (such as weights) of the classifier [11].

**Test set**: A set of examples used only to assess the performance (generalization) of a fully-specified classifier [11].

**Translating the Tweets Scores to “Positive” and “Negative” Labels**

To eliminate any bias in the scores assigned to the tweets, we assigned every tweet an average of the total score it received from three distinct labelers. This average score was rounded to the closest integer score which was in the range of 0 to 4. All tweets with an average rounded score of 1 and 2 were categorized as low/no-risk tweets or “Negative” HIV risk tweets. Similarly, all tweets with an average rounded score of 3 and 4 were categorized as high/very-high risk tweets or “Positive” HIV risk tweets. Tweets with an average score of 0 were dropped since they fell in the “Do not understand” category.

With this form of translation we observed that, from the 2000 labeled tweets, 346 were labeled as “Positive” HIV risk, 1649 were labeled as “Negative” HIV risk, and 5 as
“Do not understand”. Thus, a total of 1995 tweets were used to train the classifier. This implied that after the two stage filtering process about 17.34% of the filtered tweets were conceived as actual HIV risk tweets and a staggering 82.65% tweets were false positives.

While splitting the data into the training and testing sets, we created an 80:20 split wherein the data with “Positive” and “Negative” labels were balanced proportionately to avoid any form of skew. By focusing on classification accuracy only, we could attain an accuracy of as high as 82.65% by marking all tweets as “Negative” owing to the skew. This would essentially mean loss of all the signal in the data. Hence, it was important not only to focus on attaining a high classification accuracy, but to also attain a higher ratio of true positives to false positives.

3.3 Experiments and Results

3.3.1 Naive Bayes Classification

Method Description: Naive Bayes (NB) methods belong to the class of supervised learning algorithms. They are based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features. Given a class variable $y$ and a dependent feature vector $x_1 \ldots x_n$, Bayes theorem states the following relationship:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

For all $i$, the assumption of independence between every pair of features yields the following.

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i \mid y)$$

$$\therefore P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$
Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Rightarrow \hat{y} = \arg \max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

where $\hat{y}$ is the predicted class label for the data represented by $x_1 \ldots x_n$.

Naive Bayes classifiers have been shown to work quite well in many real-world situations such as document classification and spam filtering despite the apparently oversimplified assumptions. They require a small amount of training data to estimate the necessary parameters [12][17].

**Bernoulli Naive Bayes**

This method implements the Naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions. There could be multiple features but each one of them is assumed to be a binary-valued variable. Therefore, this class requires samples to be represented as binary-valued feature vectors.

The decision rule for Bernoulli Naive Bayes is based on:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

We modified our document vectors such that every feature was represented by a binary-valued variable based on the occurrence of the feature. This approach differs from the Multinomial Naive Bayes rule (see below). Bernoulli Naive Bayes explicitly penalizes the non-occurrence of a feature $i$ that is an indicator for class $y$, whereas the Multinomial Naive Bayes method would simply ignore a non-occurring feature [12].
Gaussian Naive Bayes

In this model the likelihood of the features is assumed to be Gaussian [12].

\[
P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)
\]

Multinomial Naive Bayes

This is variant that is often used in text classification tasks where the data are typically represented as word vector counts, like in our case. The distribution is parametrized as \( \theta_y = (\theta_{y1}, \ldots, \theta_{yn}) \) for each class \( y \), where \( n \) is the number of features (the size of the vocabulary) and \( \theta_{yi} \) is the probability \( P(x_i \mid y) \) of feature \( i \) appearing in a sample belonging to class \( y \).

The parameters \( \theta_y \) is estimated using maximum likelihood as follows:

\[
\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}
\]

where \( N_{yi} = \sum_{x \in T} x_i \) is the number of times feature \( i \) appears in a sample of class \( y \) in the training set \( T \), and \( N_y = \sum_{i=1}^{\mid T \mid} N_{yi} \) is the total count of all features for class \( y \) [12].

Using Naive Bayes Classifier for our Classification Problem

Table 3.1 summarizes the classification accuracy obtained with different Naive Bayes (NB) approaches i.e. Bernoulli Naive Bayes, Gaussian Naive Bayes, and Multinomial Naive Bayes. These learning algorithms are applied to the dataset with different features i.e. Feature Sets 1, 2, 3, 4 that are described in the section above. The testing accuracy is averaged over a 5 fold 80:20 cross-validation. These results indicate that the Multinomial Naive Bayes approach does the best in terms of classification accuracy with 84.71% testing accuracy using unigrams + bigrams + trigrams + riskwords, and a true positives
to false positives ratio of 17.78. This means that for every 17-18 true positive samples that we let into our corpus, we would have to deal with about 1 false positive sample.

Table 3.1. Naive Bayes Classification Accuracy for HIV at-risk Tweets

<table>
<thead>
<tr>
<th>Accuracy (%)</th>
<th>Bernoulli NB</th>
<th>Gaussian NB</th>
<th>Multinomial NB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>T/F</td>
</tr>
<tr>
<td>Unigrams</td>
<td>86</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Unigrams +</td>
<td>85.45</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Bigrams</td>
<td>87.50</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Unigrams +</td>
<td>85.07</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Bigrams +</td>
<td>87.50</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Trigrams</td>
<td>85.07</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Unigrams +</td>
<td>85.07</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Bigrams +</td>
<td>87.50</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>Trigrams +</td>
<td>85.07</td>
<td>82.5</td>
<td>0/0</td>
</tr>
<tr>
<td>RiskWords</td>
<td>85.07</td>
<td>82.5</td>
<td>0/0</td>
</tr>
</tbody>
</table>

Note: T/F stands for True-Positive Rate / False-Positive Rate or TPR/FPR.

### 3.3.2 Logistic Regression Based Classification

**Method Description:** Logistic Regression is a mathematical modeling approach that can be used to describe the relationship between several independent variables $X_1 \ldots X_k$ to a dichotomous dependent variable $Y$ [1]. This mathematical model is based on the logistic function that is shown in Fig 3.4. The logistic function, denoted as $f(z)$, is given by $\frac{1}{1+e^{-z}}$. As the graph indicates, the value of $f(z)$ is always between 0 and 1. Thus, this model is designed to describe a probability that is always between 0 and 1. In our case, this probability can help decipher the risk of a tweet showing HIV at-risk behavior.
Figure 3.4. Logistic Function

The logistic model is derived from the logistic function as follows:

\[ z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \]

Here, \( X_s \) are the independent variables of interest and \( \alpha \) and \( \beta_i \) are constant unknown parameters. \( z \) is essentially an index that combines the \( X_s \).

\[
 f(z) = \frac{1}{1 + e^{-z}} \to \frac{1}{1 + e^{-(\alpha + \sum \beta_i)}}
\]

Let \( Y \) be the dichotomous dependent variable that can take on values 0 and 1. The probability statement can be expressed as:

\[
 P(Y = 1 \mid X_1 \ldots X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i)}}
\]
Using Logistic Regression Classifier for our Classification Problem

Table 3.2 summarizes the classification accuracy obtained with the Logistic Regression Classifier. The classifier is applied to the dataset with different features i.e. Feature Sets 1, 2, 3, 4 that are described in the section above. The testing accuracy is averaged over a 5 fold 80:20 cross-validation. These results indicate that the Logistic Regression Classification approach does the best in terms of classification accuracy with 85.86% testing accuracy using unigrams + bigrams + trigrams + riskwords, and a true positives to false positives ratio of 18.04. This again means that for about every 18 true positive samples that we let into our corpus, we would have to deal with about 1 false positive sample.

Table 3.2. Logistic Regression Classification Accuracy for HIV at-risk Tweets

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Logistic Regression Accuracy (%)</th>
<th>Train</th>
<th>Test</th>
<th>T/F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>96.56</td>
<td>84.91</td>
<td>8.386</td>
<td></td>
</tr>
<tr>
<td>Unigrams + Bigrams</td>
<td>99.44</td>
<td>85.32</td>
<td>13.66</td>
<td></td>
</tr>
<tr>
<td>Unigrams + Bigrams + Trigrams</td>
<td>99.75</td>
<td>85.66</td>
<td>16.06</td>
<td></td>
</tr>
<tr>
<td>Unigrams + Bigrams + Trigrams + RiskWords</td>
<td>99.72</td>
<td>85.86</td>
<td>18.04</td>
<td></td>
</tr>
</tbody>
</table>

3.3.3 Support Vector Machine (SVM) Based Classification

Research has been done to analyze the particular properties of learning with text data and why SVMs are appropriate for this task. SVM’s ability to learn can be
independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. This helps us generalize even in the case of high dimensional data as long as it is separable with a wide margin using functions from the hypothesis space [6]. This is an important consideration in our case given that our feature space (with unigrams + bigrams + trigrams + riskwords) has a very high dimensionality of the order of $36K$ features.

**How SVM Works:**

The goal of a support vector machine is to find the optimal separating hyperplane which maximizes the margin of the training data [8].

The separating hyperplane that separates a $D$ dimensional space into two half-spaces is represented as follows:

$$w^T x + b = 0$$

where $w$ is the normal vector that is orthogonal to any vector lying along the hyperplane, and $b$ is the hyperparameter or the bias term.

Maximization of the margin for the hyperplane can be viewed as an optimization problem whose primal is stated as:

$$\min ||w||^2 + C \sum \varepsilon$$

$$s.t. (w.x + b)y \geq 1 - \varepsilon$$

$$\varepsilon \geq 0$$

where $\varepsilon$ is the slack parameter.
On obtaining the vector \( w \) and hyperparameter \( b \), we apply the following prediction rule for binary classification:

\[
h(x) = \text{sign}(w \cdot x + b)
\]

where \( x \) is the feature vector, \( w \) is the weights vector normal to the hyperplane, \( b \) is the hyperparameter, and \( h(x) \) corresponds to class prediction.

**Why SVMs Work Well for Text Classification**

In his paper titled “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, Thorsten Joachims outlines why SVMs work well for text classification. These reasons are described below [6]:

**High dimensional input space**: While learning text classifiers, one has to deal with very high dimensional (about 10K or more) feature space. With the curse of high dimensional data, there is always a chance of overfitting while learning classifiers. However, since SVMs use overfitting protection (which does not necessarily depend on the number of features), they have the potential to handle these large feature spaces.

The goal of a support vector machine is to find the optimal separating hyperplane which maximizes the margin of the training data.

**A few irrelevant features**: One way to avoid high dimensional input spaces is to assume that most of the features are irrelevant. Feature selection tries to determine these irrelevant features. However, in text categorization there are very few irrelevant features. Several research experiments performed suggest that even features ranked the lowest contain relevant information when it comes to text categorizations and are not completely redundant. Hence, a good classifier needs to combine several features and learn a dense concept. Aggressive feature selection may result in loss of information.
**Document vectors are sparse:** For each document, as in our case, where every document is a tweet, the corresponding document vector contains only few entries which are not zero. Kivinen et al. [7] give theoretical and empirical evidence that SVMs are well suited for problems with dense concepts and sparse instances.

**Most text categorization problems are linearly separable:** Thorsten Joachims suggests in his research [6] that text based data is linearly separable.

We use these arguments as theoretical evidence that SVMs should be a good fit for text based classification.

**Using SVM Classifier for our Classification Problem:**

Table 3.3 summarizes the classification accuracy obtained with the SVM Classifier with a linear kernel. The classifier is applied to the dataset with different features i.e. Feature Sets 1, 2, 3, 4 that are described in the section above. The testing accuracy is averaged over a 5 fold 80:20 cross-validation. These results indicate that the SVM Classification approach does the best in terms of classification accuracy with 85.91% testing accuracy using unigrams + bigrams + trigrams + riskwords, and a true positives to false positives ratio of 10.95.

**Table 3.3. SVM Classification Accuracy for HIV at-risk Tweets**

<table>
<thead>
<tr>
<th></th>
<th>SVM - Linear Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy (%)</strong></td>
<td><strong>Train</strong></td>
</tr>
<tr>
<td>Unigrams</td>
<td>99.63</td>
</tr>
<tr>
<td>Unigrams + Bigrams</td>
<td>99.94</td>
</tr>
<tr>
<td>Unigrams + Bigrams + Trigrams</td>
<td>99.98</td>
</tr>
<tr>
<td>Unigrams + Bigrams + Trigrams + RiskWords</td>
<td>99.97</td>
</tr>
</tbody>
</table>
3.4 SVM: Tuning the Hyperparameter

Our goal was not only to improve the classification accuracy, but also to improve the ratio of “true positives” to “false positives”. We performed Receiver Operating Characteristic (ROC) curve analysis to tune the hyperparameter to help us achieve our goal.

3.4.1 ROC Curve Overview

The tradeoff between obtaining more true positives at the expense of additional false positives for different values of a tuning parameter is visualized in an ROC curve by plotting this tradeoff for every possible threshold. The X-axis represents the false positive ratio, “FPR” and the Y-axis represents the true positive ratio, “TPR”. A model is first built from a set of training data and then evaluated against a set of test data, often within a cross-validation process. All ROC curves start in the bottom left corner, representing a threshold at which all cases are classified as negative. They end in the upper right corner, representing a threshold at which all cases are classified as positive. Better curves are closer to the upper-left corner (if one curve is above another at a given point on the horizontal axis, the higher curve is better at detecting true positives, while generating the same percentage of false positives as the lower curve). Curves should also always be above the diagonal, as a curve below the diagonal indicates that a model is generating more false positives than true positives.

Thus, the information presented in an ROC curve can help choose an appropriate threshold for the hyperparameter [2].

For the purpose of our experiment, $b$ is varied and the tradeoff between the TPR and FPR is calculated for every threshold. Fig. 3.5 shows the ROC Curve for tweet classification with a linear SVM with 5-fold cross-validation. Each of the 5 different
color-coded curves represents the ROC curves for 5 different test sets picked as part of the cross-validation process. This figure suggests that the hyperparameter $b$ can be tuned to a value to yield an overall expected test accuracy of 85.6% with an expected TPR/FPR value of $\sim 17$.

Figure 3.5. ROC Curve for Text Based HIV Risk Classification with SVM

This section has demonstrated how the various supervised learning based classifiers perform in terms of text classification accuracy and the ratio of true positives to false positives. In the next section we talk about how this information can be used to predict whether a user is at risk for HIV, as well as the type of risk based on their tweets. We continue to use the SVM classifier for the tweets’ risk classification.
3.5 Classifying Users for HIV risk based on their Tweets’ Risk

So far we have focused on classifying tweets for HIV risk based on their text content. In this section, we discuss how this categorization of tweets can be used to infer whether a user is at risk for HIV.

For user based HIV risk classification, we first classify all the tweets using the SVM classifier with a linear kernel that is discussed in the previous section. The hyperparameter is tuned using ROC curve analysis so that a high TPR to FPR ratio is attained. Every tweet is assigned a risk-score and a risk-label on being run through this classifier.

Let the risk-score for tweet $i$ be denoted as $\hat{Z}_t^i$. For every tweet $i$,

$$\hat{Z}_t^i = (W.X_i + b)$$

where $W$ is the weight vector learned by means of the SVM classifier, and $b$ is the tuned hyperparameter. Let the risk-label for tweet $i$ be denoted as $Z_t^i$. For every tweet $i$,

$$Z_t^i = \text{sign}(\hat{Z}_t^i)$$

where

$$Z_t^i = \begin{cases} 
1 & \Rightarrow \text{Positive HIV Risk Tweet} \\
-1 & \Rightarrow \text{Negative HIV Risk Tweet}
\end{cases}$$
We define a user’s risk-score as follows:
Let the risk-score for user \( u_j \) be denoted by \( Z_{u_j} \). For every user \( u_j \), the risk-score is the average of the sum of risk-scores of all tweets from user \( u_j \) in our HIV corpus.

\[
Z_{u_j} = \frac{1}{\text{count}(i)} \sum_{i \text{ is a tweet by } u_j} Z_i
\]

Let the risk-label for user \( u_j \) be denoted by \( Z_{u_j} \). For every user \( u_j \),

\[
Z_{u_j} = \text{sign}(Z_{u_j})
\]

where

\[
Z_{u_j} = \begin{cases} 
1 & \Rightarrow \text{Positive HIV Risk User} \\
-1 & \Rightarrow \text{Negative HIV Risk User}
\end{cases}
\]

3.5.1 User Risk Analysis with our Labeled Data

The 2000 tweets in the labeled data set were owned by 1348 distinct users. We computed each of these user’s risk-score and risk-label as defined above. We observed that the average accuracy of user risk prediction on the training set was 94.55% whereas on the test set the average accuracy was 82.84%.

3.5.2 Risk Prediction with the Global Corpus

HIV Risk Prediction for Tweets in the Global Corpus:

As described in Chapter 2, our system continuously ingests tweets from San Diego County and filters them based on the presence of risk terms as potentially HIV
at-risk tweets. As of date, we have gathered 24,062,785 tweets from San Diego County of which 69,453 have been filtered as being potentially HIV at-risk tweets. We further classified these 69,453 tweets using our text-based SVM classifier to weed out the false positives. Of the 69,453 tweets, 1838 were classified as true positives for HIV risk.

We plan to have domain experts go through these positively labeled tweets to give us feedback on the quality of these newly classified tweets in terms of signal to noise ratio as far as HIV risk analysis goes. However, a quick look at a randomly picked sample of ~100 tweets from these positively classified tweets suggests a much higher signal to noise ratio than obtained with merely filtering tweets based on risk terms. And a quick look at a random sample of ~100 tweets from the negatively classified tweets also suggests that most of these were falsely filtered as at-risk tweets and hence they ought to be discarded to lower the noise in the gathered data for risk analysis.

**HIV Risk Prediction for Users in the Global Corpus:**

For understanding the nature of risk a user is at based on the data gathered in the global corpus, we decided to observe the evolution of a user’s risk type along a temporal dimension. We developed a visualization that helped understand a particular user’s risk type (Drug, Sex, SexVenue, Homosexual, STI) over a range of months.\(^1\) Fig. 3.6 illustrates this visualization for a period of one month from our data set.

Every node represents an at-risk user as predicted by our text-based classification model. The edges between these nodes represent user connections (we discuss these in Chapter 4). The colors of the nodes are used to indicate the risk type (*Red: Drug, Green: Homosexual, Purple: Sex, Gray: SexVenue, Pink: STI*). A user’s risk type is determined by taking the majority of the risk types each of his tweets belong to in our corpus. However, it might be useful to allow visualizing multiple risk types for a single user.

\(^1\)This visualization was developed by Ajay Mohan, one of our PIRC-NET team members.
Figure 3.6. HIV risk visualization for users based on the risk type over one month

user. It would also help to associate a degree of risk for each user based on probability estimates from our binary classification based risk prediction.

Such a visualization that can help identify HIV risk categories based on risk type and the degree of risk along a temporal dimension could potentially be a useful aid for targeted interventions in the future.

This concludes our exploration of text-based classification approaches for categorizing tweets and users for HIV risk. The next chapter will explore another key
component of potential risk: users’ network connectivity. We believe that online social networks can provide insights to understand the HIV at-risk nature of groups of people.
Chapter 4
Exploring the HIV at-risk Network

The social graph of at-risk users and tweets that results from the processed tweets is rich with information. A platform like Twitter provides a lot more information that can be extracted and be used to augment our analysis. So far we have only explored a small chunk of the entire social-graph; specifically, the user-tweet connections. Fig. 2.2 describes the high-level structure of the graph that was created to store the network based information of potentially at-risk individuals, their tweets, and user-user connections. This basic network can be utilized to derive a whole new range of more meaningful user-user connections. In this chapter we explore various social relationships that we can infer from the umbrella HIV risk network shown in Fig. 2.2 for a potentially better assessment of HIV risk.

4.1 User-User Connections Derived from the HIV Risk Network

One of the key motivators of the study was a question raised by the clinicians on our team when analyzing HIV risk behavior: “Can Twitter’s network data be used as a tool to infer the social network of individuals that are at high risk of acquisition or transmission of HIV?” This led us to further explore types of social relationships on Twitter that could possibly be good indicators of real life user-user connections. After
brainstorming with the domain experts, we decided to explore the following four types of social relationships derived from our HIV risk network that indicate strong user-user connections:

1. **Conversations**: With this social relationship, we aimed to model user-user connections between users that engaged in conversations with one another via exchanging tweets which could potentially be a good indicator of them being connected in the real world. Every pair of users that engages in a direct exchange of tweets is said to have had a “conversation”.

2. **Geographic + Temporal Co-location**: Given the nature of HIV transmission, physical contact between users who are at risk is an important factor to consider. It could make them and their networks more susceptible to the risk of HIV transmission/acquisition. With the social relationship based on geographic and temporal co-location, we aimed to model user-user connections between users that were co-present in a spatial+temporal sense which could potentially be a good indicator of them being connected in the physical world. Every pair of users that were within a mile apart from one another in a time window of 1 hour are said to have been geographically and temporally co-present.

3. **Mentions**: Twitter allows the “mention” of an individual in a tweet as a mechanism to directly address that tweet to the mentioned individual. This social relationship could once again be a good indicator of user-user connections in the real world.

4. **Follows**: Twitter allows users to “follow” other individuals as a mechanism for the followers to consume content generated by the followees. We use this social relationship to model physical world user-user connections; if a pair of users follow each other then there is a higher likelihood of them being connected in the physical world.
The following subsections outline how each of these user-user connections are derived from the HIV risk network, as well as an analysis of the properties of the sub-networks derived from these connections and their visualizations.

4.1.1 Conversations

We define a "conversation" as a direct exchange of tweets between two or more users. We derive this social relationship using two basic properties from the umbrella HIV risk network built in Chapter 2: “TWEETED” and “IS_REPLY_FOR”.

The following cypher query\(^1\) illustrates how to obtain a pair of users, \(u_1\) and \(u_2\), such that \(u_1\)'s tweet, \(t_1\), is a reply to \(u_2\)'s tweet, \(t_2\).

\[\text{Cypher Query: MATCH (u1:USER) - [[:TWEETED]] - (t1:TWEET) - [[:IS_REPLY_FOR]] - (t2:TWEET) - [[:TWEETED]] - (u2:USER)}\]

We derive four attributes for each such conversation between user \(u_1\) and \(u_2\):

1. **Frequency of Conversation**: This property captures the number of unique conversations between \(u_1\) and \(u_2\). It can help gauge how often \(u_1\) and \(u_2\) tend to exchange tweets with one another.

2. **Average Length of Conversation**: This property captures the average length of continued conversations between \(u_1\) and \(u_2\). It can help gauge whether \(u_1\) and \(u_2\) tend to engage in short or longer conversations.

3. **Risk Category of Conversation**: This property categorizes the conversations into one of the 5 risk categories: Drug, Sex, Sex Venues, Homosexual, STI based on the text content of the tweet exchanges that are part of that conversation. It can help

\(^1\)Cypher is the query language of our graph database Neo\(4\)J
understand the nature of the risk that is implied by the conversation between u1 and u2.

4. **Tweets in the Conversation Chain:** This additional property allows us to actually view the text used in the conversation chain.

A graph visualization that emphasizes only the user-user conversation connections within the HIV risk network was created with these derived properties. Every node in this graph visualization stands for a user in the HIV risk network. “Conversation” edges join every pair of users said to have been involved in a conversation as per our definition. Each of these edges have the above mentioned properties. The edges are directed, the source node being the initiator of the conversation.

The graph visualization for user conversations is shown in Fig.4.1. As we can observe, there are several tightly connected clusters dispersed throughout the graph. Such clusters could be identified and analyzed separately for some additional information for HIV risk analysis. This visualization highlights the power of observing the user-user connections as a network for a more wholesome picture of the risk patterns within a network. The following cypher query is used to visualize the network from the derived database:

---
**Cypher Query:**
```
MATCH (u1:USER) - [r:DRUG|HOMO|SEX|SEXVENUE|STI] -> (u2:USER)
RETURN r
```
---

We can also visualize specific types of conversation sub-networks within the conversations network in order to focus on a certain risk category of conversations
Figure 4.1. Graph Visualization of Conversations within the HIV Risk Network

and individuals. Fig.4.2 shows the graph visualization of conversations that are of “Homosexual” nature. We can observe that this graph has a different set of tightly connected clusters than those observed in Fig.4.1. Thus, analyzing the network based on the risk category can help uncover various different clusters within the network to aid targeted intervention based on the type of risk involved.

The following cypher query is used to visualize the network based on homosexual
conversations from the derived database:

Cypher Query: MATCH (u1:USER) - [r:HOMO] - (u2:USER)
RETURN r

Figure 4.2. Graph Visualization of Homosexual Conversations within the HIV Risk Network
Fig. 4.3 shows a zoomed in peek at the user-user conversations in the network. The green nodes indicate the user nodes. They are identified by a system-generated identifier to protect the user’s anonymity. The edges between these nodes indicated conversations. They are color-coded and labeled with the risk category that the conversation belongs to.

![Diagram of conversations network](image)

**Figure 4.3.** Zoomed View: Graph Visualization of Conversations within the HIV Risk Network

Just like the visualizations based on risk category, one can observe the conversations network based on the frequency of the conversation between two users. The following cypher query shows how to view user-user conversations where the number of conversations between two users is more than 5:
Cypher Query: MATCH (u1:USER) - [r:DRUG|HOMO|SEX|SEXVENUE|STI] -> (u2:USER) WHERE r.freq > 5
RETURN r

Similarly, one can observe the conversations network based on the average length of these conversations. The following cypher query shows how to view user-user conversations where the average length of conversations between two users is more than 5:

Cypher Query: MATCH (u1:USER) -[r:DRUG|HOMO|SEX|SEXVENUE|STI] -> (u2:USER) WHERE r.avg_len > 5
RETURN r

In a graph or a network, the degree of a node is the number of edges incident on it. In other words, it is the number of edges or connections it has to other nodes in the network. The degree of a node gives us a sense of its connectivity within the network. Although the graphs are directed in nature, we consider them to be undirected for the frequency distribution of the node degrees within the network.

Fig. 4.4 shows the frequency distribution of degrees for the conversations network. We observe that while most of the nodes have a degree of 1 or 2, a considerable number of nodes have degree in the range of 3-10. Degrees of more than 10 are sparse. This implies a somewhat sparse set of user connections in terms of conversations as we have modeled with very few highly connected nodes that lead to tight clusters.
4.1.2 Geographic + Temporal Co-location

For the purpose of our study, we first define the following terms:

**Geographic Co-location**

Every tweet we gather has a geographic identification. While some of them are identifiable by geo-tags, some of them can be identified by a geographic bounding-box. When we talk about geographic co-location, we are interested in identifying individuals potentially at risk that may have been present in about the same area. The way we choose to check if two users $u$ and $v$ are geographically co-located is by computing the distance between their tweets using the tweets’ geographical co-ordinates. Users $u$ and $v$ are said to be “geographically co-located” if and only if their tweets, $t_u$ and $t_v$, are within a 1 mile radius of one another.
Temporal Co-location

Every tweet we gather also has a ‘created_at’ time-stamp. When we talk about temporal co-location, we are interested in identifying individuals potentially at risk that have tweeted within a short time-span. The way we choose to check if two users \( u \) and \( v \) are temporally co-located is by computing the time difference between their tweets using the tweets’ created_at time-stamps. Users \( u \) and \( v \) are said to have been “temporally co-located” if and only if their tweets, \( t_u \) and \( t_v \), are within a time-span of 1 hour.

Based on these definitions, “Geographic+Temporal Co-location” helps identify individuals that were within a mile apart from each other and within a time-span of an hour. We abbreviate this connection as simply “co-location”. Establishing co-location from an individual’s social footprint can help draw a direct connection from their social connections to their real world connections. These connections can also help identify potential risk-venues where a lot of at-risk individuals are identified as co-located.

A graph visualization that emphasizes only the user-user connections based on co-location within the HIV risk network was created. Every node in this graph visualization stands for a user in the HIV risk network. “COLOCATION_WITH” edges join every pair of users said to have been co-located as per our definition. Each of these edges has assigned to it a frequency property to indicate how often the users connected by that edge were co-located. The edges are bidirectional indicating a symmetric connection.

Fig. 4.5 shows the graph visualization of user-user connections based on co-location derived using Neo4j. We can observe that several clusters are interspersed within this network. Their nature varies from that observed in the conversations network.

The following query is used to visualize the co-location sub-network from the derived database:

**Cypher Query:**

```
MATCH (u1: USER) - [:COLOCATED_WITH] - (u2: USER)
```
Figure 4.5. Graph Visualization of User-User Connections based on Co-location

Fig. 4.6 zooms into this visualization to show one such cluster. Each purple node here represents a user and is identified by a system generated integer to protect the anonymity of the users. The edges represent the “colocation_with” connections as
**Figure 4.6.** Zoomed View: Graph Visualization of User-User Connections based on Co-location

explained previously.

Fig. 4.7 shows the frequency distribution of degrees for the co-location network. The symmetric nature of this network implies that the minimum degree of any node is 2.

We observe that while most of the nodes have a degree of 2, a few have degrees in the range of 4-6. Degrees of more than 10 are extremely sparse. This implies a sparser set of user connections in terms of co-location as we have modeled with very few highly connected nodes that would lead to tight clusters. The cluster sizes are also quite small as we can observe. Although, this network is sparser it has valuable information than can be directly correlated with the physical world.

### 4.1.3 Mentions

Twitter allows a user to directly address other users via tweets using their Twitter handles. E.g. “@alice Let’s party tonight!” is a way to address this tweet directly to Alice. Social relationships based on user mentions could be a good indicator of real world user-user connections since they indicate direct communication between two or
Figure 4.7. Frequency Distribution of Node Degrees within the Co-location Network

more users. This information is obtained from the umbrella HIV risk network derived in Chapter 2 using two basic properties: “TWEETED” and “MENTIONED_IN”

The following cypher query illustrates how to obtain a list of users $u$ and $v$ such that $v$ mentions $u$.

\[
\text{Cypher Query: MATCH (u:USER) - [:MENTIONED_IN] -> (:TWEET) <- [:TWEETED] - (v:USER) RETURN u.id_str, v.id_str}
\]

The “mentions” relationship between two users has the following property:

**Frequency**: It indicates the number of times a certain user has mentioned a certain other user in their tweets. For example, if user $v$ mentions user $u$ 5 times, the “mentions” edge
from \( v \) to \( u \) will have a frequency property with value 5.

A graph visualization that emphasizes only the user-user connections based on user mentions within the HIV risk network was thus created. Every node in this graph visualization stands for a user in the HIV risk network. “MENTIONS” edges join every pair of users where one of them mentions the other. Each of these edges have the above mentioned ‘frequency’ property. Also, these edges are directed, the source node representing the user mentioning the user represented by the destination node.

Fig. 4.8 shows the graph visualization obtained from the umbrella HIV risk network based on user mentions using Neo4j.

This graph shows several clusters that differ in nature from those observed in the networks based on conversations and co-location and has been obtained using the following query on the derived database.

```
Cypher Query: MATCH (u:USER) - [r:MENTIONS] - (v:USER)
RETURN r
```

Fig. 4.9 zooms into this visualization. Each blue node here represents a user and is identified by a system generated integer to protect the anonymity of the users. The edges represent the “mentions” connections as explained previously.

Fig. 4.10 shows the frequency distribution of degrees for the mentions network. We observe that a large number of nodes have a degree of 1, while a considerable number of nodes have degrees in the range of 2-5. Fewer nodes have degrees in the range of 6-10. Nodes with degrees of more than 10 are sparse. This implies that the mentions network is less sparsely connected overall as compared to the conversations and co-location networks, but there are still very sparse densely connected clusters within the network.
Figure 4.8. Graph Visualization of User-User Connections based on User Mentions in Tweets

4.1.4 Follows

Twitter allows users to “follow” other individuals as a mechanism for the followers to subscribe to the updates or consume the content generated by their followees. Since users tend to follow those whom they either connect with in real life or whose opinions resonate with their own, the follower/followees sub-network within the HIV risk network can provide some useful cues towards understanding real-world user connections.

The following cypher query illustrates how to obtain a list of users $u$ and $v$ such
Figure 4.9. Zoomed View: Graph Visualization of User-User Connections based on User Mentions in Tweets

Figure 4.10. Frequency Distribution of Node Degrees within the Mentions Network

that $u$ follows $v$. 
Cypher Query: MATCH (u:USER) - [r:FOLLOWS] ->(v:USER) 
RETURN r

A graph visualization that emphasizes only the user-user connections based on the follows relationship within the HIV risk network was thus created using Neo4j. Every node in this graph visualization stands for a user in the HIV risk network. “FOLLOWS” edges join every pair of users where one of them “follows” the other. Also, these edges are directed, the source node representing the follower and the destination node representing the followee. Fig. 4.11 shows the graph visualization obtained from the umbrella HIV risk network based on “follows”.

The previous graph is obtained using the following query on the derived database.

Cypher Query: MATCH (u:USER) - [r:FOLLOWS] ->(v:USER) 
RETURN r

This graph shows several clusters that can each be individually studied to understand user connections based on the “follows” relationship. While the previous networks mostly had disjoint clusters, this network has an interestingly unique form of clusters where most of the dense clusters are interconnected amongst themselves.

Sometimes it might be useful to look at bidirectional “follow” edges for a stronger connection between two users in the real world. The following query shows how this could be achieved:

Cypher Query: MATCH (u:USER) - [r1:FOLLOWS] ->(v:USER), 
MATCH (v) - [r2:FOLLOWS] ->(u) 
RETURN r1, r2
Fig. 4.12 zooms into this visualization. Each yellow node here represents a user and is identified by a system generated integer to protect the anonymity of the users. The edges represent the “follows” connections as explained previously.

Fig. 4.13 shows the frequency distribution of degrees for the follows network. We observe that most of the nodes have a degree of 1, while a considerable number of nodes have degrees in the range of 2-5. Fewer nodes have degrees in the range of 6-10.
Figure 4.12. Zoomed View: Graph Visualization of User-User Connections based on Follows

Nodes with degrees of more than 10 are sparse. This implies that the follows network is also sparsely connected overall with some tight clusters.

Given these observations, our next step is to understand if our current knowledge allows us to answer our original question by combining text and network based approaches. In other words “Can we improve our HIV risk analysis by augmenting our text-based analysis with the topological information that we have derived?”. We explore this in the next chapter.
Figure 4.13. Frequency Distribution of Node Degrees within the Follows Network
Chapter 5

Augmenting Text-based HIV Risk Categorization with Network Information

In Chapter 3, we discussed a purely text-based approach for HIV risk categorization for tweets and users. In Chapter 4, we learned about the different network structures and social relationships that emerge from the network of tweets and tweeters. In this chapter, we explore how information about these social relationships can be used to augment the text-based categorization approach to improve the HIV risk analysis. The main motivation here is that social media footprint is a good indicator of a person’s real life behavior, and users connected in the social graph are generally more likely to have similar inclinations [14]. This could potentially help extrapolate that a user’s risk for HIV acquisition or transmission could be influenced by their direct or indirect connections within the network. Our hope is to have a proof of concept that indicates that incorporating social-network information can lead to improvements in HIV risk analysis over the merely text-based approach discussed in Chapter 3.

5.1 Premise

Our previous approach for HIV risk analysis, just like several other existing analysis approaches with social networks, assumes that texts are independent and identically
distributed (i.i.d.) and focuses on building a sophisticated feature space to handle noisy and short messages. Speaking of Twitter specifically, this approach fails to take advantage of the fact that the tweets are networked data which may contain useful semantic clues that are not available in purely text-based methods. Social science findings suggest that sentiment consistency and emotional contagion are observed in social networks. Research in sentiment analysis with social networks based on these findings has shown better performance in handling the short and noisy data [4]. Two social processes, selection and influence, are proposed to explain this phenomenon: people befriend others who are similar to them (Homophily), or they become more similar to their friends over time (Social Influence). Both these explanations suggest that connected individuals are more likely to have similar behaviors or hold similar opinions [4]. Based off of these findings, we propose a simple machine learning model that uses social relationship information in the context of HIV risk analysis.

5.2 Proposed Model

Let the set of tweets being examined for HIV risk be $T$, and the set of users whose tweets are being examined for HIV risk be $U$. Our goal is to determine which of these users exhibit “Positive” HIV risk and which of them exhibit “Negative” HIV risk. For each user $u_i \in U$, we have a set of tweets $t_{ui}$. We also know which users $u_j$ are connected to $u_i$ in the social graph with one of the following connections: conversation, co-location, mentions, follows. Fig. 5.1 shows such a directed social graph instance with the users, their tweets, and their connection information. This incorporates both textual as well as social-network information in a single graph where the nodes correspond to either the tweets or the users, and the connecting edges depict the relationship between a user and a tweet or between two users.
We formally define this social graph $G$ as: $G = \{ U \cup \{ t_u | u \in U, t_u \in T \}, E \}$. The edge set $E$ is the union of two sets: the tweet edges $\{(u_i, t_{u_i}) | u_i \in U, t_{u_i} \in T \}$ implying that $u_i$ posted tweet $t_{u_i}$, and the network induced user-user edges [14]. Edge $(u_i, u_j) \in E$ can be of one of the following types: Conversations, Co-location, Mentions, and Follows.

We propose a mixture model that combines the text information with the network information which results in a new hybrid model. First, we define two matrices that capture the user-tweet and user-user connections. Let $M$ be the matrix that contains the user-tweet connections. Every row of $M$ corresponds to a user $u_i \in U$ and every column corresponds to a tweet $t_{u_i}$ such that $u_i \in U, t_{u_i} \in T$. $m(i)$ corresponds to the total number of tweets by user $u_i$. 

Figure 5.1. Social Network Structure with Tweets and Users
Similarly, we define a connection matrix \( C \) that contains the network induced user-user connection information.

\[
M_{i\alpha} = \begin{cases} 
1 & \text{if } u_i \text{ tweets } t_\alpha \\
0 & \text{otherwise}
\end{cases}
\]

Every element \( C_{ij} \) in matrix \( C \) denotes if users \( u_i \) and \( u_j \) where \( u_i, u_j \in U \) are connected. We constrain the connection matrix, \( C \), to be symmetric in nature for the simplicity of the model. This means that if \((u_i, u_j) \in E\) then \((u_j, u_i) \in E\) holds as well.

Let \( X_{i\alpha} \) correspond to the text-only feature vector that corresponds to tweet \( t_\alpha \) tweeted by user \( u_i \). This feature vector is formed from the unigrams + bigrams + trigrams + riskwords as discussed in Chapter 3. Let \( W \) be the weight vector and \( b \) be the hyperparameter derived using the SVM classifier with a linear kernel that was discussed in Chapter 3.

Tweet \( t_\alpha \)'s HIV risk score from text-only features is calculated as \( W.X_{i\alpha} + b \). The HIV risk score for user \( u_i \) with text-only information is thus the average of the HIV risk scores for every tweet \( t_\alpha \) made by user \( u_i \) (\( t_{i\alpha} \)). Let us denote the HIV risk score of a user \( u_i \) using text only as \( \hat{Z}_i \).

\[
\hat{Z}_i = \frac{1}{m(i)} \sum_{\alpha=1}^{\lfloor M_i \rfloor} (W.X_{i\alpha} + b)
\]
Let $Z_i$ denote the HIV risk for user $u_i$ that is to be inferred from our network-based model, and $\gamma$ be the mixture co-efficient such that $0 < \gamma \leq 1$. We use $\text{sign}(Z_i)$ as the risk label for user $u_i$.

$$\text{sign}(Z_i) = \begin{cases} 
1 & \Rightarrow \text{Positive HIV Risk User} \\
-1 & \Rightarrow \text{Negative HIV Risk User}
\end{cases}$$

The objective function for our prediction model based on network is as follows:

$$\mathcal{L} = \gamma \sum_i (Z_i - \hat{Z}_i)^2 + (1 - \gamma) \sum_{ij} C_{ij}(Z_i - Z_j)^2$$

We want to minimize this objective function with respect to $Z_i$. Observe that this is a convex function. Hence, we take the gradient of this function and set it to zero, and solve for $Z_i$ such that this objective function is minimized. Let $D$ be the diagonal matrix obtained by summing the rows of $C$ such that every diagonal element corresponds to the sum of the corresponding row in $C$. Let $L = D - C$. Since $C$ is symmetric in nature, we observe that: $\sum_{ij} C_{ij}(Z_i - Z_j)^2 = Z^T LZ$

$$\therefore \frac{\delta \mathcal{L}}{\delta Z_i} = 0 \Rightarrow \gamma(Z_i - \hat{Z}_i) + (1 - \gamma)(LZ)_i = 0$$

$$\therefore [(1 - \gamma)L + \gamma I]Z = \gamma \hat{Z}$$

$$\therefore Z = ((1 - \gamma)L + \gamma I)^{-1}(\gamma \hat{Z})$$

We thus derive $Z$, the HIV risk prediction vector for users based on our new network-based model.
5.3 Experiments

Let us now understand how the different types of user-user connections within the HIV risk network help inform our risk analysis based on the new network-based model that we have derived in the previous section. We perform our network-model based predictions with the labeled dataset that was obtained as explained in Chapter 2 to measure the impact of the network connections on the prediction accuracy. For each of the user-user connections explored in Chapter 4, we construct a user-user connection matrix. Let us call this connection matrix $C$. Since there are 1348 users in our labeled dataset, $C$ is a $1348 \times 1348$ symmetric matrix. The mixture coefficient, $\gamma$ is also varied from $0^+$ to 1, $0^+$ meaning that the model uses only network features, and 1 meaning that the model uses only text features and no network features. We already know from Chapter 3 that with text-only features the user risk prediction accuracy is 94.55% with the training data and 82.84% with the test data. The training and testing accuracy values for text-based user prediction are represented by horizontal lines in the graphs in the following subsections just to show the baseline. Viewing the training and testing accuracy values derived from the network-based model with respect to this baseline would help us understand the interplay between the text and network features for user risk prediction.

The following subsections go over how this connections matrix $C$ is constructed for each of the connection types, as well as its impact on the user risk prediction accuracy.

5.3.1 Conversations

In Chapter 4, we define a conversation between two users as an exchange of tweets between them. Fig. 4.1 shows the graph visualization of user-user conversations within the HIV risk network. For our new network-based prediction model, we initially modeled $C$ based on direct conversations only.
Direct Conversations:

\[ C_{ij} = \begin{cases} 
1 & \text{if } u_i \text{ and } u_j \text{ have a conversation} \\
0 & \text{otherwise} 
\end{cases} \]

Fig. 5.2 shows the training and testing accuracy for user-user connections based on direct conversations. We can observe that for \( \gamma \) in the range 0.3 to 0.5 the training accuracy is 94.47% and the testing accuracy improves to 83.14% from 82.84% that was originally obtained with the text-based classification. This shows that augmenting the text based classification with direct conversation connections provides some additional signal.

The connection matrix obtained for direct conversations was sparse. Hence, we also modeled \( C \) such that it was based not only on direct conversations but also on connections between users that were reachable from one another via the direct conversation links. We call such conversations indirect conversations. Thus, the corresponding element \( C_{ij} \) for users \( u_i \) and \( u_j \) who are reachable form one another is weighted by the inverse of the minimum number of hops along the direct conversation edges between \( u_i \) and \( u_j \).

Indirect Conversations:

\[ C_{ij} = \begin{cases} 
\frac{1}{\text{minimum hops}(u_i,u_j)} & \text{if } u_i \text{ can reach } u_j \text{ via undirected conversation edges} \\
0 & \text{otherwise} 
\end{cases} \]

The following cypher query illustrates how to obtain the minimum number of hops between users \( u \) and \( v \) if \( u \) has indirect conversations with \( v \).

**Cypher Query:** MATCH p = shortestPath ((u:USER {id_str:'u.id_str'})-[*]-v:USER {id_str:'v.id_str'})
Fig. 5.2. Classification Accuracy with Network Features: Direct Conversations

\[
(v:USER \{id\textunderscore str:`v.id\textunderscore str`\})) \text{ WHERE } \text{LENGTH}(p) \neq 0 \text{ RETURN } u.id\textunderscore str, v.id\textunderscore str, \text{LENGTH}(p)
\]

Fig. 5.3 shows the training and testing accuracy for user-user connections based on indirect conversations. We can observe that for $\gamma$ in range 0.5 to 0.6 the training accuracy is 94.47% and the testing accuracy is 83.14%. Thus, having a less sparse connection matrix also does not help in improving the user risk prediction accuracy much.

5.3.2 Co-location

In Chapter 4, we define two users to be co-located if they were within a mile apart from each other within a time-span of one hour. Fig. 4.5 shows the graph visualization of user-user connections based on co-location. To model the connection matrix $C$ we make
use of both direct and indirect co-location as done previously in the case of conversations in order to have a less sparse matrix. For direct co-location, the corresponding element $C_{ij}$ for users $i$ and $j$ is 1 if $i$ has been co-located with $j$.

**Direct Co-location:**

$$C_{ij} = \begin{cases} 
1 & \text{if } u_i \text{ and } u_j \text{ are co-located} \\
0 & \text{otherwise} 
\end{cases}$$

For indirect co-location, the corresponding element $C_{ij}$ for users $i$ and $j$ who are reachable form one another is weighted by the inverse of the minimum number of hops along the direct co-location edges between $i$ and $j$.

**Figure 5.3.** Classification Accuracy with Network Features: Indirect Conversations
Figure 5.4. Classification Accuracy with Network Features: Direct Colocation

**Indirect Co-location:**

\[
C_{ij} = \begin{cases} 
\frac{1}{\text{minimum hops}(u_i, u_j)} & \text{if } u_i \text{ can reach } u_j \text{ via undirected co-location edges} \\
0 & \text{otherwise}
\end{cases}
\]

Fig. 5.4 shows the training and testing accuracy for user-user connections based on direct co-location. We can observe that for \(\gamma\) in range 0.3 to 0.5 the training accuracy is 94.12% and the testing accuracy is 83.14%.

We do not observe any improvement in the user risk prediction accuracy with user connections based on indirect co-location. However, on combining user-connections based on direct conversations and direct co-location we observe that for \(\gamma\) in range 0.6 to 0.7 the training accuracy is 94.39% and the testing accuracy is 83.14%. Fig. 5.5 shows the training and testing accuracy for user-user connections obtained on combining direct
Figure 5.5. Classification Accuracy with Network Features: Direct Colocation and Direct Conversation

conversations and direct co-location.

5.3.3 Mentions

As described in Chapter 4, we say a user $u$ mentions user $v$ if $u$ directly addresses $v$ via tweets using $v$’s Twitter handle. Fig. 4.8 shows the graph visualization of user-user connections based on mentions. To model the connection matrix $C$ based on mentions we make use of both direct and indirect mentions once again. For direct mentions, the corresponding element $C_{ij}$ for users $u_i$ and $u_j$ is 1 if $u_i$ has mentioned $u_j$ (and vice versa). The symmetry of $C$ is maintained.
**Direct Mentions:**

\[
C_{ij} = \begin{cases} 
1 & \text{if } u_i \text{ mentions } u_j \text{ or if } u_j \text{ mentions } u_i \\
0 & \text{otherwise}
\end{cases}
\]

For indirect mentions, the corresponding element \( C_{ij} \) for users \( i \) and \( j \) who are reachable form one another is weighted by the inverse of the minimum number of hops along the direct mentions edges between \( i \) and \( j \).

**Indirect Mentions:**

\[
C_{ij} = \begin{cases} 
\frac{1}{\text{#minimum hops}(u_i,u_j)} & \text{if } u_i \text{ can reach } u_j \text{ via undirected mentions edges} \\
0 & \text{otherwise}
\end{cases}
\]

With the connections based on direct “mentions”, we observe that for \( \gamma \) in the range 0.6 to 0.7 the user prediction training accuracy is 93.39\% and the test accuracy is 83.43\% whereas for \( \gamma \) in the range 0.8 to 0.9 the user prediction training accuracy is 94.47\% and the test accuracy is 83.14\%. Fig. 5.6 shows the training and testing accuracy for user-user connections based on direct mentions.

On combining direct mentions with the previously derived connections i.e. direct conversations and co-location we observe that for \( \gamma \) between 0.4 and 0.6 the training accuracy is 94.04\% and the test accuracy is 83.43\% whereas for \( \gamma = 0.8 \) the training accuracy is 94.30\% and the test accuracy is 83.43\%. Fig. 5.7 shows the training and testing accuracy for user-user connections based on direct mentions combined with direct conversation and co-location information. The overall test accuracy is between 83.14\% and 83.43\% for \( 0<\gamma<1 \). These connections show marginal improvement in the user risk prediction accuracy as far as the test data goes over a larger range of \( \gamma \)s.
5.3.4 Follows

As described in Chapter 4, we capture user-user connections based on the “follows” social relationship. Fig. 4.11 shows the graph visualization obtained from the umbrella HIV risk network based on “follows”. To model the connection matrix $C$ based on “follows” we make use of both direct and indirect follows once again. For direct follows, the corresponding element $C_{ij}$ for users $u_i$ and $u_j$ is 1 if $u_i$ follows $u_j$ or if $u_j$ follows $u_i$. The symmetry of $C$ is maintained.

**Direct Follows:**

$$C_{ij} = \begin{cases} 
1 & \text{if } u_i \text{ follows } u_j \text{ or } u_j \text{ follows } u_i \\
0 & \text{otherwise}
\end{cases}$$

![Figure 5.6. Classification Accuracy with Network Features: Direct Mentions](image)
For indirect follows, the corresponding element $C_{ij}$ for users $u_i$ and $u_j$ who are reachable from one another is weighted by the inverse of the minimum number of hops along the direct follows edges between $u_i$ and $u_j$.

**Indirect Follows:**

$$C_{ij} = \begin{cases} 
\frac{1}{\text{minimum hops}(u_i,u_j)} & \text{if } u_i \text{ can reach } u_j \text{ via undirected follows edges} \\
0 & \text{otherwise}
\end{cases}$$

With the connections based on direct “follows”, we observe that for $\gamma$ in the range 0 to 0.2 the user prediction training accuracy is 94.56% and the test accuracy is 83.14% whereas for $\gamma$ in the range 0.3 to 0.4 the user prediction training accuracy is 94.65% and the test accuracy is 83.14%. Fig. 5.8 shows the training and testing accuracy for user-user connections based on direct follows.
On combining user connections from follows with the connections derived previously i.e. direct conversations, co-location, and mentions we observe that for $\gamma$ in the range 0 to 0.4 the training accuracy lies between 92.98% and 93.70% and the test accuracy improves to 83.72%. For $\gamma$ in the range 0.5 to 0.6 the training accuracy lies between 93.78% and 94.04% and the test accuracy is 83.43%. And for $\gamma$ in the range 0.7 to 0.9 the training accuracy lies between 94.21% and 94.47% and the test accuracy lies between 83.14% and 83.43%. Fig. 5.9 shows the training and testing accuracy for user-user connections based on direct follows along with direct conversations, co-location, and mentions.

For all the user connections we have explored so far, the user prediction accuracy improves a little on the test data but does not improve much on the training data with our network-based prediction model.
Due to small prediction accuracy improvements observed with our network-based model, we tried to understand if incorporating user connections had much signal at all. In order to check this, we initialized the connection matrix $C$ with random positive integers and then ran our prediction model.

Fig. 5.10 shows the training and testing accuracy values using randomly initialized connections in $C$. This uncovers an interesting observation. Both the training and testing user risk prediction accuracy values drop to 80.52% for $\gamma < 1$. This shows that the network connections do impact the risk prediction accuracy values. The reason for little improvement with the connections derived from the social relationships based on conversations, co-location, mentions, and follows could be due to the sparseness in these
user-user connections as well as their noisy nature.

Figure 5.10. Classification Accuracy with Network Features: Random Connections

5.4.1 Comparing Network Only with Text Only Analysis

Overall in our analysis, we ideally would like to observe two specific categories of people at risk:

- Users who are at risk due of their at-risk behavior
- Users who are at risk because of their network that puts them at risk

We use text-based user risk prediction as a measure of a user’s risk score due to their at-risk behavior. A user’s risk score due to their network is computed as the average of the risk score of the direct neighbors. Users that had no direct neighbors were dropped from this analysis. Fig. 5.11 shows how every user’s risk score based only on the network co-varies with that based only on the text from their tweets.
The top left quadrant of Fig.5.11 represents those users who are at risk not due to their own behavior but due to the network they are associated with which is potentially at risk. This is useful information that we were able to uncover thanks to our network-based model. Such information could not have been obtained while analyzing individuals for risk without looking at their network. This could potentially be useful to aid targeted intervention measures.

This concludes our exploration of the network-based classification methods for categorizing users for HIV risk. It might be worth working towards reducing the sparseness as well as the noise in the user connections for better user risk prediction using the network-based model. It might also help to look at some other forms of user connections that may be good indicators of risk influence as well as to work towards refining the prediction model for further improvement.
Chapter 6

Inferences and Directions for Future Work

PIRCNET is a step towards progressing from the traditional intervention measures towards targeted, evidence-based intervention measures. We have demonstrated how information from digital social networks, specifically Twitter, can be harnessed to understand the nature of HIV risk for users and their connections. In particular, we extrapolate a user’s risk based on their own risk-behavior using the text-based learning methods as well as the risk imparted to them by virtue of being connected to an at-risk network. Experimental results show us that the text-based risk prediction as well as network-based risk predictions can potentially help to better categorize the at-risk communities based on social media footprint.

6.1 Summary of Inferences

The key motivation for this research is the need for targeted intervention measures to effectively use resources at hand to fight the spread of HIV. Keeping this in mind, the main goal was to try and understand how to better characterize and identify vulnerable populations using tweets, social relationships, and related information inferred from Twitter’s social graph. PIRCNET provides the base infrastructure to gather and analyze
tweets whose users may potentially be considered at risk. In addition to gathering tweets, the infrastructure also gathers supplementary information related to the users and relevant network information. To effectively deal with the large corpus of gathered data, a text based classification approach was first used to improve the filtering process for tweets so as to eliminate noise in the form of false positives. Owing to previous empirical evidence of SVM being a good fit for text based classification for documents that are sparse and noisy, we use an SVM based supervised learning approach to train the text-based classifier that helps predict the at-risk tweets as well as users.

The text-based classification method weeds out a large number of false positives thereby improving the signal to noise ratio by a large extent. It is useful to not only identify at-risk communities, but also at-risk venues based on the activities that would put people at HIV risk. Owing to the networked nature of user connections on a platform such as Twitter, online social relationships can potentially map to real world connections among users from spatial as well as temporal dimensions. In particular, we look at relationships such as conversations, co-location, mentions, and follows since they could potentially be good indicators of physical world connections among people. Our network based model for user risk prediction suggests that user connections modeled from online social relationships can provide some additional signal to help improve the prediction accuracy. Although we did not observe a significant improvement by using the network model with our data, our model demonstrates a novel approach that takes into account specific context around text-based classification and is therefore a promising direction for further exploration.

Our inferences about at-risk communities and locations over a period of time can help inform interactive visualizations that domain experts can interact with to make real time decisions about targeted intervention.
6.2 Challenges

Along the way there have been several challenges, both technical as well as social. The first and foremost challenge has been improving the signal in the gathered data and also verifying if it has indeed improved. We rely on labels assigned to the filtered at-risk tweets by domain experts while training our classifier for risk prediction. However, since the labeling process is based on the labeler’s instincts for the most part, this is inherently a subjective process. Also, within the community, the domain experts’ familiarity with the risk words and related vocabulary may differ. Such factors could lead to discrepancies in the labels obtained. Also, because we could only crowdsource the labeling task within the community of domain experts, it was difficult to obtain a larger set of labeled tweets as and when required for the purposes of training and testing.

Finding a pattern from tweets has more challenges than natural language processing because the vocabulary keeps evolving, there is significant non-standard localization, and no grammar rules as such. As time progresses, the nature of the slang used for risk words is bound to change. This flux in the vocabulary needs to be addressed by continuously updating the risk words in our filter.

Further, we hoped to superimpose the risk network derived from the real world to the one derived from our analysis from online social media data to understand the similarities and differences. For this reason, we tried to recruit patients that underwent HIV diagnoses at the Hillcrest’s AVRC Center with their consent for which they would agree to share their Twitter handles and would also be willing to be contacted by our research team for any further inquiries. These were potentially HIV positive or HIV at-risk individuals due to the fact that they underwent tests for HIV diagnosis. However, this ended up being a big social challenge owing to the reluctance of patients to opt into the study. We had only 19 patients who have undergone HIV diagnoses sign up for our
study. The small numbers made it harder for us to draw any significant inferences based on their data.

Our network-based prediction model was designed for symmetric user connections only. However, some user connections such as “follows” are inherently asymmetric. Hence, it is important for the model to capture such unidirectional connections while making risk predictions. Our network-based prediction model is also symmetric in terms of the risk influence imparted based on a user’s connections. While a non-risk user being highly connected to several at-risk users is considered to be at risk by the model, we do not want a high-risk user being highly connected to several non-risk users to diminish his riskiness. The simplicity of our model fails to address this.

Apart from this, one of the reasons why the network-based prediction model fails to improve the user risk prediction accuracy significantly is that the connections we use at the moment may be noisy. We only take into account the existence of the connections and not their frequencies. For instance, in the case of conversations, it might be worthwhile to weight the connections between users with the frequency of the conversations. Also, since the user connections from the data gathered might be sparse, it might be useful to come up with other such forms of derived connections that are good indicators of user connections in the real world. Refining the design of our network based prediction model to address these issues could potentially help improving the user risk prediction.

With the number of tweets gathered and the massive sizes of the HIV risk networks that are derived, the algorithms for fetching related network information and risk prediction take a rather long time (∼ 20 hours) as of date. We would need to design the data collection infrastructure as well as our algorithms such that they scale well with the enormous amounts of data that will continue to be gathered.

Other challenges are related to intervention ethics. Based on our inferences, a suitable and ethical means of intervention that is non-intrusive needs to be devised at a
community or spatial level, and not at a personal level. Translating our inferences into suitable intervention action items is of key importance for its effectiveness.

6.3 The Road Ahead

Our foray into using machine learning to support HIV intervention schemes opens up a myriad of interesting directions for future work. Since improving the signal in the data needs to be an iterative, feedback driven process, it would be useful to adopt a reinforcement learning based approach to keep refining the text and network prediction models. We could have domain experts provide continuous feedback regarding the signal in the data we gather for HIV risk analysis in terms of the ratio of true positives to false positives to help continue refining the models. A possibility currently being looked at by our team is to have HIV test patients hand label the potentially at-risk tweets filtered by our system in the waiting areas at the clinic. This could help us procure a lot more labeled samples for more robust training and testing purposes.

Now that we know that the network-based model helps uncover useful information related to user-risk, it would be useful to design the model to eliminate symmetry of risk influence as explained in the previous section. It would also be useful to eliminate the constraint of symmetry in the user connections to allow asymmetric connections. Weighing the user connections with frequency of occurrence, span of occurrence or other such qualifying parameters might be useful in establishing the strength of the user connections. Also, in the current implementation we used a purely text-based supervised learning approach to train our classifiers, without accounting for any network-based features during the training process. It might be useful to consider the network parameters while training the classifier along with the text-based features. This might also help improve the user risk analysis.
To deal with and draw inferences from the massive amounts of data gathered, we may need to consider adopting a MapReduce based model that perform tasks in parallel on large data sets.

Also, it would be useful to develop interactive visualizations to serve as an effective public health tool that exploits the spatial and temporal aspects of the data gathered and classified. To understand the similarities and differences in the HIV risk behaviors between the online world and the physical world, it is important to find other, more effective means to recruit patients from the real world for our study.

In conclusion, this is a small yet significant step in being able to effectively check the spread of HIV while also enabling the at-risk communities to contribute by making them a part of the process. We believe that it is possible to build an effective HIV risk surveillance radar that exploits the approach put forward in this thesis, and we are hopeful that such a surveillance radar would help plan effective targeted intervention schemes.
# Appendix A

## HIV Risk Words

**Table A.1. HIV Risk Words**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Homosexual</th>
<th>Sex</th>
<th>Sex Venue</th>
<th>STI</th>
</tr>
</thead>
<tbody>
<tr>
<td>methamphetamine</td>
<td>homo</td>
<td>bareback</td>
<td>aqua day spa</td>
<td>full house</td>
</tr>
<tr>
<td>meth</td>
<td>queer</td>
<td>barebacking</td>
<td>vulcan sauna</td>
<td>chlamydia</td>
</tr>
<tr>
<td>ice</td>
<td>gay</td>
<td>creampie</td>
<td>the hole</td>
<td>clap</td>
</tr>
<tr>
<td>speed</td>
<td>homosexual</td>
<td>raw dogging</td>
<td>t room</td>
<td>gonorrhea</td>
</tr>
<tr>
<td>cocaine</td>
<td>queen</td>
<td>basket shopping</td>
<td>tea room</td>
<td>syphilis</td>
</tr>
<tr>
<td>coke</td>
<td>tink</td>
<td>cottaging</td>
<td>groovy laidback</td>
<td>syf</td>
</tr>
<tr>
<td>crystal</td>
<td>bear</td>
<td>lucky pierre</td>
<td>baja betty’s</td>
<td>sif</td>
</tr>
<tr>
<td>crank</td>
<td>chicken</td>
<td>on the make</td>
<td>brass rail</td>
<td>syph</td>
</tr>
<tr>
<td>blow</td>
<td>chicken hawk</td>
<td>rumpy-rumpy</td>
<td>bourbon street</td>
<td>pox</td>
</tr>
<tr>
<td>tina</td>
<td>donald duck</td>
<td>zipper club</td>
<td>caliph lounge</td>
<td>bad blood</td>
</tr>
<tr>
<td>crack</td>
<td>girl scout</td>
<td>yard boy</td>
<td>cheers</td>
<td>vd</td>
</tr>
<tr>
<td>tweak</td>
<td>Grimm’s fairy</td>
<td>brass rail</td>
<td>flicks</td>
<td>dose</td>
</tr>
<tr>
<td>tweaker</td>
<td>smurf</td>
<td>ring snatcher</td>
<td>the loft</td>
<td>gleet</td>
</tr>
<tr>
<td>dopez</td>
<td>vampire</td>
<td>Pratt</td>
<td>ma4</td>
<td>morning drip</td>
</tr>
<tr>
<td>glass</td>
<td>wolf</td>
<td>lucky pierre</td>
<td>numbers</td>
<td>running rage</td>
</tr>
<tr>
<td>flailing</td>
<td></td>
<td></td>
<td>pecs</td>
<td>drip</td>
</tr>
<tr>
<td>speeding</td>
<td></td>
<td></td>
<td>redwing</td>
<td>white</td>
</tr>
<tr>
<td>booty bump</td>
<td></td>
<td></td>
<td>rich’s</td>
<td>racehorse</td>
</tr>
<tr>
<td>booty bumping</td>
<td></td>
<td></td>
<td>eagle</td>
<td>the clam</td>
</tr>
<tr>
<td>plug</td>
<td></td>
<td></td>
<td>sro</td>
<td>clam chowder</td>
</tr>
<tr>
<td>plugging</td>
<td></td>
<td></td>
<td>urban mo’s</td>
<td>gooey stuff</td>
</tr>
<tr>
<td>butt rocket</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bumping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


