UNIVERSITY OF CALIFORNIA SAN DIEGO

A Data Science Approach for Real-Time HIV-Risk Analysis on Twitter

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

in

Computer Science

by

Amarnath Raju Vysyaraju

Committee in charge:

Professor Nadir Weibel, Chair
Professor Susan Janet Little
Professor Lawrence K Saul

2019

The Thesis of Amarnath Raju Vysyaraju is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

Dedicated to my kind and loving Mother, Father, Brother, and all my great Teachers and Well-wishers.
I am who I am because of you.

EPIGRAPH

Errors using inadequate data are much less than those using no data at all.

Charles Babbage

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It had been a remarkable experience working on this project and writing this Thesis. I am extremely grateful to the committee chair Prof. Nadir Weibel for being such a marvelous advisor. If it had not been for all the brainstorming sessions and discussion meetings, this work would not have been possible. Thank you for helping me gain a worldly view in the field of research while offering different perspectives in approaching the problems.

A special thanks to Prof. Lawrence Saul and Prof. Susan Little for being on the committee and lending their immense knowledge to evaluate the project's progress and its future.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of work, study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

ABSTRACT OF THE THESIS

A Data Science Approach
for Real-Time HIV-Risk Analysis on Twitter

by

Amarnath Raju Vysyaraju

Master of Science in Computer Science

University of California San Diego, 2019

Professor Nadir Weibel, Chair

HIV is still a major epidemic and although significant progress in treatment has been achieved, a functional cure for HIV is still far away, and a great deal of effort is currently focused on the prevention of HIV. Prevalence of HIV has recently prompted clinicians and public health officials to take a look at social media as the source of digital epidemiology. This thesis introduces our data science approach aimed at capturing HIV-related trends based on multidimensional data from Twitter. We show how our platform can help clinicians understand people's risk behavior, and ultimately guide in HIV prevention. Our design is flexible and extensible, and currently employs a collection of techniques that span crowd-sourcing, natural

language processing, image classification, supervised machine learning, and graph data analysis to classify at-risk tweets and user groups. In our experiments, we have established the relationship between an individual user's risk along with the network's risk for HIV based on their actions on Twitter. This infrastructure will serve as a foundation for building visualizations and real-time analytical tools for studying the prevalence of HIV-risk to better inform prevention resources.

Chapter 1

Introduction

1.1 Motivation

The Human Immuno-deficiency Virus (HIV), when left untreated, can lead to irreparable damages to the immune system, ultimately resulting in the death of the person who carries this virus. As of 2018, HIV has claimed more than 35.4 million lives worldwide and 36.9 million people globally were living with HIV in 2017, with an estimated 1.8 million individuals becoming infected every year [1]. The US Center for Disease Control and Prevention (CDC), reports that 1.1 million live with AIDS in the USA and 40,000 become newly infected every year [2].

Although no large-scale cure exists yet for HIV, this does not mean that HIV treatment cannot be administered; nowadays HIV-infected people can live almost normal lives if infections are detected early on, HIV treatment is started as soon as possible, and the medications' regimen is continuous. Many efforts in public health are therefore aimed at prevention and early detection, with the goal to both lower the number of at-risk individuals, and identify HIV-positive individuals fast so that they can be introduced into care as early as possible.

While research and clinical care are moving fast to enable detection with HIV early testing [3]—which can detect HIV infection as early as one week after exposure—contemporary methods of HIV-risk assessment have still a long turn-around time. Risk assessment typically relies on static data from national census statistics and other surveys that dates back a few

years[1], and in some cases are complemented by local data documented by the HIV clinics about their own patient population. Most of the time, however, these data only represent the specific subset of HIV at-risk individuals that are already seen at clinics for treatment or testing, and therefore the HIV population with the highest probability of infection or transmission remains undiscovered.

Our goal is to use a computational approach, rooted in the analysis of people's online communication, to uncover this population at an earlier stage. Through a near real-time monitoring system we seek to help public health researchers to respond to HIV risk more effectively. Specifically, we focus on social media that produce real-time data like Twitter[2]—as opposed to traditional bio-surveillance, which relies on information collected at clinics—and exploit this medium to tackle health-related issues.

Other work used Twitter as a vehicle for public health interventions [4] and demonstrated the feasibility of using social networking data as a method for evaluating and detecting HIV risk behaviors [5, 6]. Our goal is to build on these efforts and exploit the opportunities that Data Science presents today for real-time processing and analysis of large quantities of data, to create a platform for HIV risk analysis on Twitter. The ultimate goal of our platform is to help clinicians prune demographic information and social connections in the local population, and initiate more targeted prevention efforts.

In this context, the contributions of this thesis can be summarized in the following areas:

- The design of our data science pipeline towards understanding online HIV-risk behavior.
- The specific text filtering techniques, in combination with computer vision, that enables better understanding of HIV-risk behavior on Twitter.
- An introduction to a graph model that enables understanding of how HIV risk networks create and evolve over time.

[1] https://aidsvu.org, https://www.cdc.gov/NCHHSTP/Atlas
[2] http://twitter.com

- The interfaces to facilitate the grounding of our data science infrastructure to identify HIV risk online through the active participation of HIV at-risk individuals in local HIV testing clinics.

- Description of a new set of HIV-risk features that are generated through our platform and how making those available to clinicians and researchers can potentially transform HIV prevention efforts.

- The use of some of these features to hypothesize and validate the Risk model that quantifies the Risk of Tweets, Users, and User networks.

1.2 Ethical Considerations

In designing and developing the data science infrastructure described in this thesis we heavily build on ethical considerations that have been previously discussed [7, 8] to ensure that data generated from a vulnerable population like HIV-positive or HIV at-risk individuals is handled in the correct way. Specifically, this work has been reviewed by the University's human subjects protection program that carefully assessed the protections we put in place to conduct both Twitter data collection, anonymization procedures, as well as the crowd-sourcing of labeled data and keywords described in the thesis below. Moreover, the output of our pipeline is always only reported as aggregated and anonymized data to prevent identification of individuals on Twitter.

Chapter 2

Related work

Before describing our platform architecture in the next chapter, we discuss below the related work, focusing on how data science techniques are being utilized to understand users' behavior. We also describe other approaches that have been introduced to mine data from social media (and specifically Twitter), and we present how others used social media to understand HIV-risk related behaviors and support HIV prevention efforts.

2.1 Data Science to Understand User Behavior

Data science is a powerful tool, currently used to understand human behavior in a variety of ways. Many of the actions performed in real life are tracked by digital media, and a great deal of research is in progress to extract significant inferences from this data. Fields such as Natural Language Processing (NLP), graph data analytics, image processing, and recommender systems, all play a key role in understanding and analyzing human information to generate tangible research and business results. Classic examples are—Linden and colleagues' use of collaborative filtering for a real-time item-to-item recommendation for Amazon.com customers [9], Youtube's usage of deep learning to recommend videos to its users [10], and Google's deep learning algorithms to address the NLP problem of generating *Smart* replies to emails in the *Inbox by Gmail* application [11]. Businesses also use NLP for the purpose of sentiment analysis to capture favorable comments on their products [12].

These techniques are not only useful for businesses, but they open up incredible opportunities that can be employed also as part of digital epidemiology [13] to better support health-care systems. Our research represents an example of such attempts, as we employ data science for public health, specifically focused on understanding HIV risk in a community to help prevent HIV.

2.2 Mining Social Media and Twitter Data

People provide a variety of information through many of the social media they utilize. Extracting appropriate information from social media is an active research area, as shown in [14] where Jin and colleagues examine the understanding of user behavior using online social networks.

In particular, given their publicly available data, micro-blogging websites like Twitter are becoming particularly interesting when it comes to understanding people's behavior. Twitter is a platform where users can post quick updates (called *Tweets*) about their life, thoughts, and feelings, usually with a limited number of characters. Twitter gives users the option to follow each other, such that a user's followers can see the other person's tweets on their homepage. Users have a provision to comment on others' tweets and be part of public online discussions. As of February 2019, Wikipedia reports about 321 million active users on Twitter.

Twitter data—in particular, user level aggregation of Twitter activity—was performed to help predict and understand users' Big Five personality traits in [15, 16]. These user-profile based studies have also been performed to find political affiliations and affinity for certain businesses, e.g. [17]. While the results of these analyses could inform tailored advertising, these studies primarily focus on a broad estimation of a user's personality. A more personalized approach would be needed to track traits that show risk towards a disease like HIV.

We believe that social media data, and Twitter, in particular, show great potential when paired with powerful data science approaches. *Natural language processing* can be used to

process and understand the user-generated text in the form of status updates, comments and messages [11]. *Image recognition and processing* can be used to understand themes and objects in images and other media shared by users [18]. *Graph data analytics* is helpful to understand and analyze the relationship between entities in social networks; [19] is an example of how Graph data analysis is used to recommend new friends in social networks.

In this thesis, we build on these ideas and introduce a data pipeline focused on capturing traits in human behavior that lead to HIV-risk. As explained in the next section, our infrastructure extends these approaches by aggregating Twitter data per user and further explore the connections between at-risk users in emerging social networks on Twitter.

2.3 Social Media for HIV

While social media represents an interesting platform to understand human behavior in general, several studies showed that this is the place where there is an abundance of conversations that can be linked to HIV-risk as well. In particular, Hospers and colleagues showed that the online environment facilitates the meeting of men with a different demographic profile, within the MSM (Men who have Sex with Men) community [20]. Similarly, the Asia Internet MSM Sex Survey in Taiwan [21] reported that 73.9% of MSM had sex with partners found online and that having sex with online partners was associated with a history of Sexually Transmitted Infections (STIs), recreational drug use, and an increasing number of sexual partners. McFarlane and colleagues showed that even early on, during the initial development of the Internet, people who sought sex online appeared to be at greater risk for STIs [22].

Therefore, social media is a potential ground for HIV-related interventions, and thus public health departments have ventured into social networking sites as venues for efficacious health education intervention [23], specifically studying the effectiveness of messages sent via Facebook[1]. Other work demonstrated how the use of iPOL (internet Popular Opinion Leaders) on Facebook was found very effective to disseminate HIV-related information to MSM [24].

[1] http://facebook.com

While these studies show that intervention can be achieved online, and on social media, they do not attempt to measure the risk towards HIV before the actual infection occurs. On the other hand, Ireland et al. [25] showed that by combining Twitter data extraction and sophisticated geocoding techniques, public health researchers could possibly implement preventive measures for a possible outbreak. Similar to what we propose in this paper, Young and colleagues ventured into the collection and filtering of tweets based on risk-terms [5]. They employ tweet filters mainly based on two categories: Drug abuse and Sex-related behaviors. Their filters do not capture tweets related to other specific behaviors that might be considered at risk as well, such as tweets related to Gay Sex or Sex Venues. They use simple risk-term based filtering schemes for tweets and perform tweet-level analysis, without however aggregating risk of multiple tweets or multiple connected users. Also, while our research focuses on a proof of concept for Twitter data collected at the county level, the presented approach could possibly be extrapolated to a state or even national level.

All in all, our platform facilitates the integration of a number of filters that include simple risk-term based filtering, and machine learning models for text classification. Our work also ventures into the thematic analysis of the media attached to the tweet. The platform we are presenting in the next section will serve as a foundation towards quantifying the risk of tweets, the risk of users, and the risk of user groups, based on users relationship among each other.

2.4 Overall Hypothesis

The CDC[2] indicates that higher risks of HIV transmission in the United States are correlated with specific groups of people. They cite higher risk in relation to (1) living in a community where many people have HIV infections, (2) having frequent bisexual or gay unprotected anal sexual intercourse, (3) being an injection drug user, (4) having another Sexually Transmitted Infection (e.g. chlamydia or gonorrhea). In addition, some specific social and ethnically di-

[2]https://www.cdc.gov/hiv/basics/transmission.html

verse communities are at higher risks; this includes African Americans, Hispanics/Latinos, and transgender man/women who have sex with men.

Single efforts have been performed to exploit the dimensions of risk outlined above. Young et al. discussed the use of Social media tools by MSM for meeting new sexual partners on the Internet [26]. They demonstrate that it is useful to understand the at-risk populations' activity on social networking platforms to be able to scale culturally tailored HIV prevention interventions. [20] is another such study that dwells into the importance of social media monitoring for at-risk behavior in MSM. These studies draw a parallel between the online activities of people at-risk, and their risk towards HIV. Buntain et al., on the other side, show how to analyze the time and location patterns of drug use by mining Twitter data [27], while Yang et al. show the importance of social media for tracking drug usage by analyzing Instagram posts and accounts [28]. Finally, HealthMap[3] is a perfect example of a real-time internet surveillance tool for epidemics.

Our work is inspired by these approaches but proposes a single data science pipeline built on these validated risk dimensions that aggregate this information towards better informing HIV risk from a multi-dimensional perspective.

Previous UC San Diego Contributions (Acknowledgements):

The work mentioned in this thesis is a continuation of the work [6, 29, 30] done by former Students Narendran Thangarajan, Purvi Desai and Ajay Mohan under the guidance of Dr. Nadir Weibel, Dr. Lawrence Saul, Dr. Amarnath Gupta, and Dr. Susan Little.

The project was taken up by Narendran, who started to explore the possibility of using Twitter for data collection and characterizing HIV risk in the Twitter population as mentioned in the paper titled "Analyzing Social Media to Characterize Local HIV At-risk Populations" [6]. The initial setup of the pipeline and the exploration of using supervised machine learning are discussed in the paper.

Purvi's Thesis [29] expands the work done previously by implementing some of the

[3] https://www.healthmap.org/en/

machine learning models and experimenting with the data. It looks at different accuracy measures compared to what has been used in this Thesis. Her thesis also formalizes the idea of having User risk scores based on their own Twitter activity and based on their connections. We have used these foundations to formulate and hypothesize the risk scores for the tweets and users in a different way as explained in Chapter 6.

Ajay's Thesis [30] introduces the work done towards the parallelized machinery that was set up for the data collection and storage pipeline which we had adapted to the needs of our work. This framework is clearly explained in Chapter 3.

We express our gratitude to all these researchers because their research serves as a premise to our work.

Chapter 3

Data Pipeline for HIV-RISK on Twitter

In order to track HIV risk behavior on Twitter, our work focuses on identifying and extracting risky behavior which usually leads to HIV in real life. Our platform builds on the previous approaches outlined above, as well as on previous work presented in [6, 7, 30]. In particular, it exploits the same HIV tweet classification framework based on five different categories of HIV Risk. In this work, we renamed and redefined the risk categories in the following way:

1. **Drug Use**: Any behavior related to the use of injection drugs like- Cocaine, Heroin or Meth.

2. **Sexually Transmitted Diseases**: Having or showing symptoms of any STD like- gonorrhea, chlamydia, genital herpes, HPV, HIV/AIDS.

3. **Gay Sex**: Men having multiple male partners, specific gay-related sex terms like- anal condom-less sex, etc.

4. **Sex Venues**: Tweets that mention places in and around San Diego, which are well-known and have a risk of HIV transmission that is potentially higher (nightclubs, bathhouses, sex venues, etc.)

5. **Sex Practices**: Other risky sexual practices, such as having condom-less sex, multiple sex partners, etc.

We call these categories *Risk buckets* and in the remainder of this thesis, we typically understand *HIV riskiness* on Twitter as showing traits that belong to any of the risk buckets above.

Our data collection infrastructure directly interfaces with Twitter and facilitates the collection of the tweets[1] using our *Collector Process* that is based on the *Tweepy* Python library [31] that implements the Twitter Streaming and Searching API. Our pipeline uses MongoDB[2] as the database since it provides ease of storage and access to tweets as semi-structured documents. As our application deals with a large amount of data, several on-demand data querying, and processing needs to be executed continuously, making a NoSQL solution a better choice.

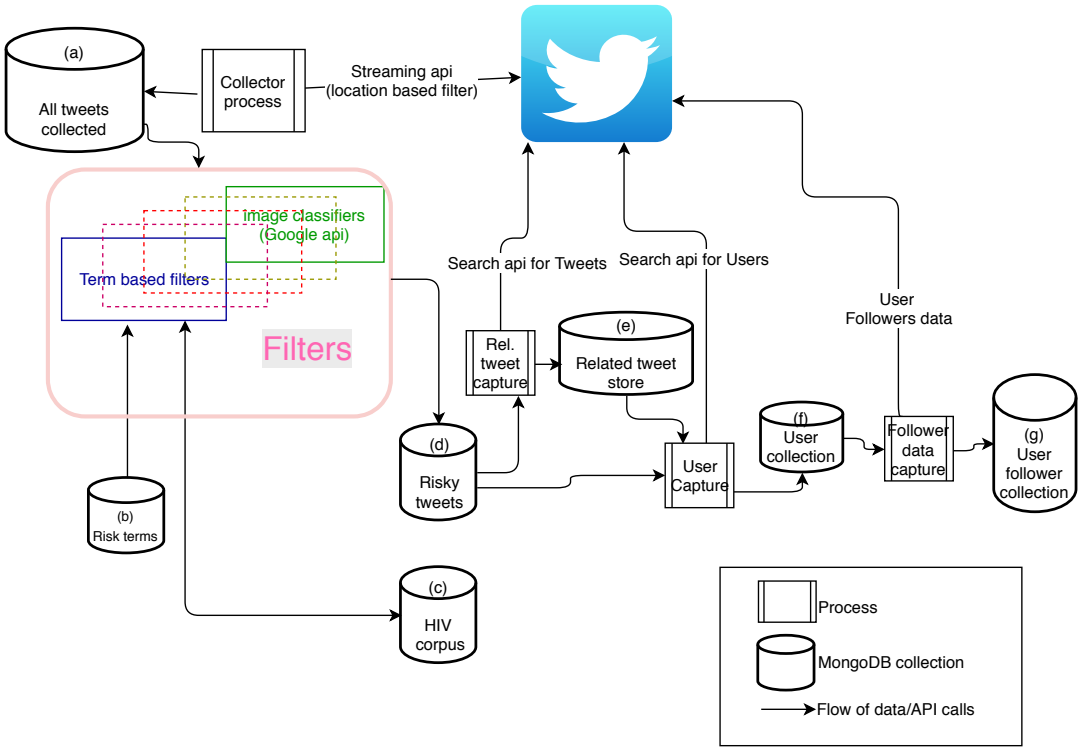

Figure 3.1. Overall Data Pipeline used to extract, filter and classify tweets for HIV risk and persist into our MongoDB collections.

[1]In order to preserve privacy, any tweet that is presented in this thesis has been paraphrased so that the content is represented truthfully, without compromising anonymity, as recommended in [8].

[2]http://mongodb.com

The data pipeline is based on a batch clustering component that comprises a number of Python scripts to enable live tweet filtering and classification. Additionally, in order to investigate relationships among users and their tweets we retrieve related information about these users and tweets from Twitter using their REST APIs. Finally, this information is fed into a number of Neo4j[3] graph database instances that help us in modeling the relationship between people and their tweets. Fig. 3.1 illustrates our overall pipeline, while Fig. 3.2 shows the Database schema for the Neo4j instances. We explain both of them in the following section.

3.1 Data Collection and Classification Pipeline

Collector process — We use the Twitter streaming API to collect our tweets corpus. Since we are interested in informing real-world interventions, we restrict tweets to only the local territory, i.e. San Diego County. We use the location-based filter by using a bounding box encompassing the surrounding region and store these raw tweets in a MongoDB collection (a) as shown in Fig. 3.1 (top-left).

Filters — We designed our platform with flexibility in mind. So, multiple HIV-risk filters can be defined, added, or removed to fit different use cases. These filters help in classifying the tweets stored in the MongoDB collection (a) into either Risky or Non-risky based on the filter-specific categories. We store the risky tweets into a different MongoDB collection (d) as shown in Fig. 3.1 (center). We will discuss the current set of filters below in this section.

Related Tweet Capture — In the next step, we consider the HIV at-risk tweets and we gather tweets that are related to them. A related tweet could mean a *retweet* or a *reply* for a risky tweet. We use the Twitter Search API based on `Tweets` to find these related tweets. We store these related tweets in a different MongoDB collection (e) (Fig. 3.1, center).

User Data Capture — In order to capture the risk behavior of the local populations, we have

[3] http://neo4j.com

to capture the Users who indicate a risk behavior towards HIV. We gather information about the users who tweeted the tweets in the collections (d) and (e) and store their data in MongoDB collection (f) (Fig. 3.1, right). We use the Twitter Search API based on `Users` to gather the User-specific information.

User Connections Capture — We gather direct connections for each user in the collection (f) by looking at their *followers* and *friends*. For instance, with two users *a* and *b*, if *b* follows *a* on Twitter, then *b* will be in the *followers* list of *a*, and *a* will be in the *friends* list of *b*. By gathering this information, we are populating our database with more users that were not captured previously via the location-based streaming collector process explained above.

3.2 HIV Term based filters

We recruited a number of *domain experts*—2x Clinicians from local HIV clinics and 4x HIV-infected patients from our University's Community Advisory Board (CAB)— who procured a list of risk terms for each of the risk buckets listed above. These risk terms are used for the first-level filtering to extract those tweets more inclined to talk about risk behavior.

Here are some examples for each of the risk buckets:

- **Drug use**: Meth, Ice, Coke, Crack
- **Sexually Transmitted Diseases**: AIDS, HIV, Gonorrhea
- **Gay Sex**: Bear, Drag Queen, Queer
- **Sex Venues**: The Loft, Bourbon Street
- **Sex Practices**: Hook up, Buggery, Bareback

3.2.1 Handling False positives

In order to handle false positive cases where tweets are marked as risky because of words like `ice`—ice is a nickname for methamphetamine, and also means the solid state of water—we created exclusion and inclusion lists that better qualify each of the HIV risk terms.

We add terms that usually co-occur with the risk term in a non-risky context to the exclusion list. For example, `ice`, when used with `cream`, is non-risky. Therefore, the exclusion list of the risk term `ice` contains the term `cream`. Those terms that co-occur with the risk term in a risky context are added to the inclusion list. For example, `ice`, when used with the word `smoke`, is usually in a risky context. Therefore, the inclusion list of the risk term `ice` contains the term `smoke`.

The risk terms for each of the risk buckets explained above are populated into a collection (b), see Fig. 3.1 (left), so that each risk term is augmented with a list of exclusion/inclusion terms as explained above: the *HIV-Risk Term based filter* populates the tweets from the collection (a) into the collection (c) (Fig. 3.1) only if the tweet satisfies any of the following conditions:

A tweet has a risk-term but does not have any of the risk-term's exclusion terms. E.g.*"Bareback feels good. But, I hate to catch an STD"*.

A tweet has a risk-term and has one or more of the risk-term's inclusion term. These tweets are populated into MongoDB collection (c) even if there is an exclusion term because the presence of an inclusion term takes precedence over the presence of an exclusion term. E.g. *"Some dude just msged me on Instagram asking me if I wanted to meet later and snort some Coke!! LMFAO"*.[4]

An example Tweet that does not satisfy these conditions from our data is: *"The world is saying that I should replace meals with ice cream #icecreamformeals"*.[5]

While we created these inclusion and exclusion lists with the assistance of our domain experts and by manually looking at sample tweets that have risk terms, this approach does not

[4] *snort* is an inclusion term for the risk-term *Coke* in our database.

[5] *cream* is an exclusion term for the risk-term *ice* in our database.

scale. We address how to extend and enhance these inclusion/exclusion lists in a dynamic way below while explaining our crowd-sourcing infrastructure.

3.2.2 Machine Learning based filters

The HIV classification filter explained above is fully deterministic and based on the HIV risk-terms and the inclusion/exclusion criteria. While it performs as prescribed, an analysis of randomly selected tweet batches showed that in some of them the number of false positives (tweets that were not at-risk but have been classified as at-risk) was high, mostly due to tweets that did not follow our inclusion/exclusion criteria list. This analysis gave us an indication that possibly the rest of the data could also show similar classification quality.

Given the large size of the corpus, we decided to design a Machine Learning (ML) filtering infrastructure as part of our data pipeline. The infrastructure is customizable and allows the integration of any ML-based classifier as a filter. To demonstrate the effectiveness of this approach, we created a simple Support Vector Machine (SVM) filter, based on a labeled set of tweets classified by a group of experts.

With the help of 30 domain experts, mostly HIV-infected patients from our University's CAB, we gathered labels for nearly 600 tweets aimed at classifying the tweet as showing a risk behavior or not. These labels allowed us to train a simple SVM model for the tweets in the collection by looking at the *Tf-Idf* of consecutive words (unigrams, bigrams, and trigrams) in the tweets in combination with our risk word. Tweets that are classified as risky by this trained SVM are currently stored in the collection (d) (Fig. 3.1, center).

The flexible infrastructure of our platform allows us to retrain the SVM model with more data for getting a better accuracy (see the crowd-sourcing section below). The labeled data is also retained and could help in experimenting with other classification models in the future.

3.2.3 Image classifiers

When analyzing some of the classified tweets we noticed that many had an image attached to them. In facts, about 10% of the tweets we captured so far have media (mostly images). In order to increase the success rate in capturing HIV-risk tweets, we equipped our platform with filters that could process the media content in the tweets. The current implementation is using Google's Cloud Vision API to capture the content of the images attached as media for a tweet. We capture the broad objects and themes of the image as text and look for the presence of risk-terms including the inclusion and exclusion criteria. The tweets that are flagged by this filter are deemed risky and populated in the collection labeled (d), Fig. 3.1 (center).

3.3 HIV Risk Network Model

In order to better investigate and visualize the network of users and their tweets, in our infrastructure, we instantiate the *risky* user and tweet data stored in MongoDB on two different Neo4j graph databases. Fig. 3.2 shows the data model representing this network of users and tweets.

Our first model (Fig. 3.2, top) represents the *direct* relationship between nodes that are derived from the data in MongoDB. Examples of these relationships are $User \Rightarrow Tweeted \Rightarrow Tweet$ and $User \Rightarrow Follows \Rightarrow User$.

The second network model (Fig. 3.2, middle) contains only $User$ nodes with both direct and derived relationships from the first graph. We use Cypher queries[6] to gather the derived relationships from the first graph and populate them in the second graph. This model contains simple relationships like $User \Rightarrow Follows \Rightarrow User$ and also complex relationships like $User \Rightarrow Drug \Rightarrow User$, which implies that the users have had a *Drug* based `risky` conversations.[7] Another example of a complex relation captured in the derived graph is a $User \Rightarrow Mentions \Rightarrow User$ relationship.

[6] https://neo4j.com/developer/cypher

[7] We define a *conversation* as an event when one user's tweet acts as a reply to another user's tweet

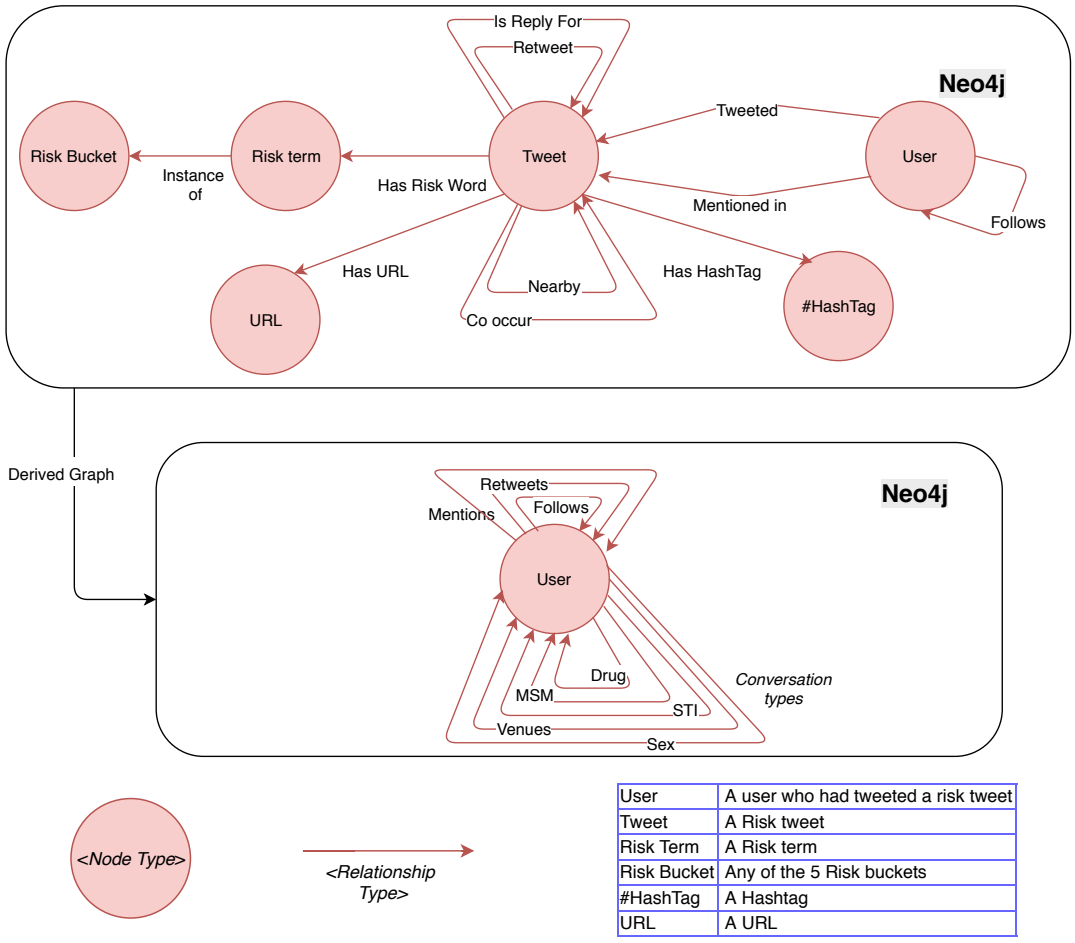

Figure 3.2. Database schema of the data in both of the Neo4j Graph Database instances: Showing all the types of nodes that are present and the relationship types that could exist between each of the node types. Note that in this figure, *Drug* = **Drug Use**, *MSM* = **Gay Sex**, *Venues* = **Sex Venues**, *Sex* = **Sex Practices**, *STI* = **Sexually Transmitted Diseases**

Fig. 3.3 shows an example of $User \Rightarrow Drug \Rightarrow User$ relationships for a subset of 500 users.

3.3.1 Cypher Queries

Cypher[8] is a graph query language that allows efficient querying and updating of property graphs.

[8] https://en.wikipedia.org/wiki/Cypher_Query_Language

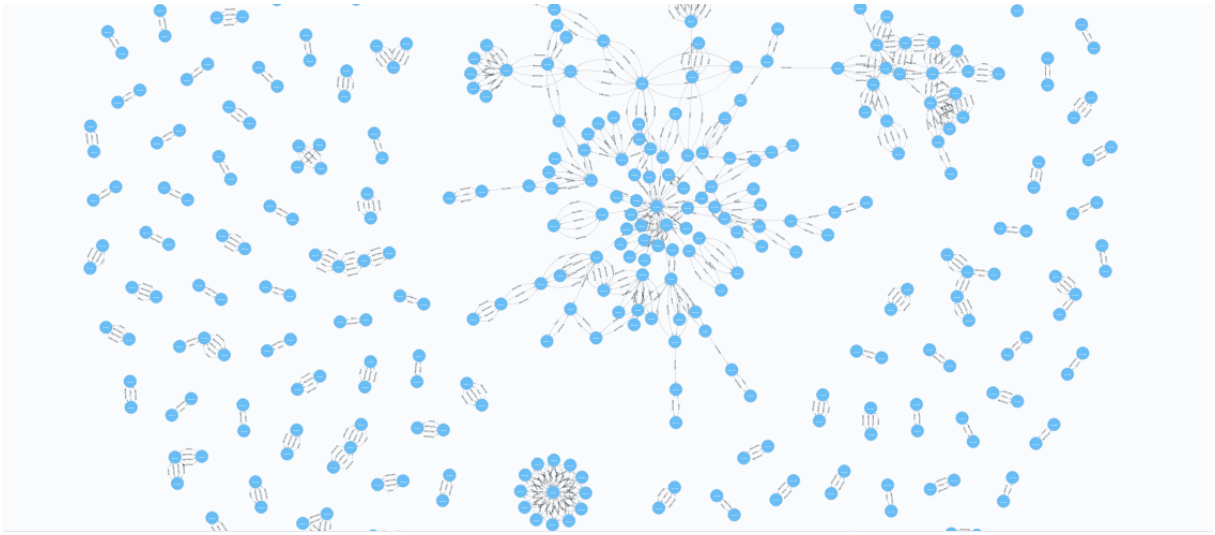

Figure 3.3. $User \Rightarrow Drug \Rightarrow User$ relationships extracted from the current Neo4j data.

The following is an example of a Cypher query that captures the conversations between users in the graph.

$$MATCH$$

$$(u1:USER)-[:TWEETED]->(t1:TWEET)$$

$$<-[:IS\_REPLY\_FOR]-$$

$$(t2:TWEET)<-[:TWEETED]-(u2:USER)$$

$$where\ u1 <> u2\ return\ u1,u2,t1,t2$$

All the $user \Rightarrow MENTIONS \Rightarrow user$ relationships can be gathered from the first graph

with the following Cypher query.

$$MATCH$$

$$(u1:USER)-[:MENTIONED\_IN]->(t:TWEET)$$

$$<-[:TWEETED]-(u2:USER)$$

$$where\ u1 <> u2\ return\ u1, u2$$

Chapter 4

Crowd-Sourcing Infrastructure

The premise of our pipeline is that we can programmatically understand when a tweet shows HIV-risk behavior and then further analyze the tweet, the related conversations, and the network of people around the involved Twitter users. Also, our hypothesis is that we will be able to understand real-world HIV-risk by looking at risk emerging from Twitter conversations and social interactions. In order for our premises to gain further validity, both of these assumptions need to be strengthened and rooted in the real world.

To address this problem, we have built an infrastructure to exploit crowd-sourcing from the local HIV community, as well as compare Twitter-behavior of individuals we *think* are at risk—the ones that our infrastructure is able to surface—with the behavior of people we *know* are at risk—people in the community that are engaging in actual risk behavior (Fig. 4.1). Our crowd-sourcing infrastructure has been designed as a tool that we are deploying to local HIV clinics and testing centers and that will be made available to all patients undergoing HIV clinical care, as well as people testing for HIV. We define these individuals as *at risk in the real world*, either because they already have HIV, or because they are frequently performing HIV testing, which in turns is a signal that they might be exposing themselves to risk.

The main premise of the crowd-sourcing tool is to leverage the knowledge of the at-risk community to crowd-source the data required to fulfill our agenda, and specifically to allow us access to the following data:

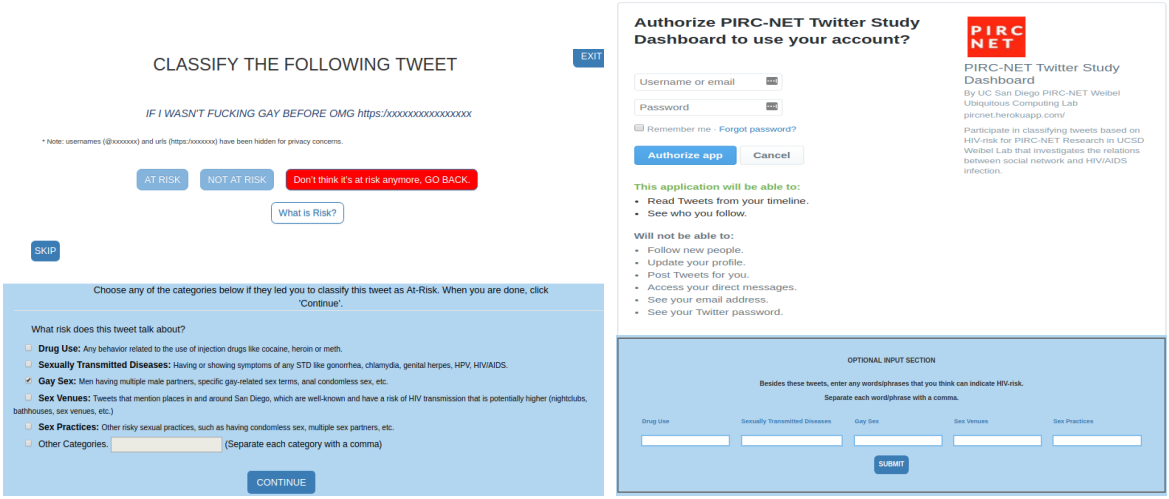

Figure 4.1. Screenshots from the crowd-sourcing dashboard. (Left) An example tweet being labeled by a participant; (Right-bottom) Provision for the participants to provide risk-related vocabulary; (Right-top) The Login page where participants agree to participate and share their Twitter data with us.

1. **Read-only Access to the Twitter accounts of the people defined as at-risk and who agreed to participate in our study** (reviewed and approved by UCSD's IRB). We gather this data when people login to our dashboard via the login page as shown in Fig. 4.1(right-top). When data from a critical number of participants will be available, the information in the participants' Twitter accounts will help us compare and contrast Twitter behavior across known and inferred HIV risk users.

2. **More tweets labeled as truthfully *Risky* or *Not-Risky***. This will allow us to retrain our ML classifications models and ultimately to improve our HIV risk-term based filters. As shown in Fig. 4.1(left), we sequentially display tweets identified as risky and ask the participants for their opinion about the real risk behavior shown in the tweet. If they label the tweet to be risky, they can assign a risk-bucket based on the tweet's context. In order to avoid annotator bias [32], each tweet is labeled by more than one participant, and a consensus through majority-vote is established.

3. **Additional risk terms from different demographics (age, ethnicity, etc.) and using potential *slang* language**. The current risk terms that are the basis of our filtering mecha-

nism have been selected by domain experts, but they do not provide enough generality to uncover many risk behaviors currently expressed in tweets. We believe that the HIV-risk population is a great source for those risk-terms, and we provide participants with the possibility to add new risk terms (see Fig. 4.1, right-bottom). Once the newly added risk terms are provided they are reviewed by our clinical and community experts before being integrated into our system.

4.1 Vocabulary Adaptation and Classifiers Remodeling

The crowd-sourcing dashboard explained above, facilitates the collection of more risk-terms, that gets added to the existing list of risk-terms after a review by our experts. Similarly, the ground-truth labels that are provided by participants are used to continuously retrain our SVM classifiers.

As explained above, finding the correct risk-words and understanding if a tweet is really showing risk-behavior is error prone. In order to support a dynamic system that has a great potential to improve, we integrated the crowd-sourcing interface directly into our data science platform, in such a way that it allows the rich labeled data-set to retrain machine learning classifiers through an automatic remodeling tool. Specifically, new risk terms are continuously added to the MongoDB collection (b) and this allows dynamic re-filtering of all the collected tweets in the collection (a) to repopulate the Risky tweets in the collection (d) of Fig. 3.1. Similarly, ML models are retrained automatically with new labeled data and provide a collection of models that can be interactively chosen to better classify the collected tweets.

This solution is not perfect, and as the word list keeps increasing indefinitely the performance of our infrastructure might be reduced significantly. Similarly, the ML model may start to over-fit, limiting the generality of the filtering task. Therefore we need to ensure that expert supervision helps in maintaining a healthy risk term database and a robust ML model. We envision clinical and community helpers to review new risky words prior to integration, and data

science experts to review automatic measures of performance of the ML models.

Acknowledgements

We are very grateful to the work done by our undergraduate students, Samarth Aggarwal, Momin Khan, Mu He, Yi-Chen Lee and Clark Phan who worked on creating, implementing and participating in the design of this crowd-sourcing dashboard.

Chapter 5

Exploratory data analysis

5.1 Experimental Data

In order to put the data-pipeline explained in Chapter 3 to valuable use, we have captured some data to experiment with and see if we could make good use of the data that we have collected for drawing inference regarding the possible trends in the risk for HIV in the Twitter population. The ultimate goal is to implement the ideas that worked on experimental data on the main data pipeline in Chapter 3

5.1.1 Users

We collected the users who tweeted or were mentioned in the Tweets that contained Risky terms from the framework described in Chapter 3. These are the tweets in the collection (c) of Fig. 3.1 (bottom). At the time, the number of users was - 147,280.

5.1.2 Tweets

We collected the tweets for the above users between *March 1, 2019* to *April 14, 2019* (UTC time). We were able to collect the Tweets for 110,189 users only because the others were protected. The total number of tweets collected were - 24,999,269.

5.1.3 User Relations

We captured the following relationships between the users as per their tweets during that time (*March 1, 2019* to *April 14, 2019*(UTC time)).

- $A \Rightarrow MENTIONED \Rightarrow B$: This relationship is created when a user A mentions user B in his/her tweets.
- $A \Rightarrow REPLIED \Rightarrow B$: This relationship is created when a user A replies to a tweet of user B.
- $A \Rightarrow RETWEETED \Rightarrow B$: This relationship is created when a user A retweets a tweet of user B
- $A \Rightarrow QUOTED \Rightarrow B$: This relationship is created when a user A quotes a tweet of user B

For each of these relationships, we are capturing the strength of the relationship by keeping the number of times such activity has occurred.

By using the above set of relationships we create a new relationship $A \Rightarrow ACTIONED \Rightarrow B$ with a parameter *count*. This relationship is created whenever there is an existing *MENTIONED / REPLIED / RETWEETED / QUOTED* relationship between users A and B. The parameter *count* for the *ACTIONED* relationship is the sum of the count parameters of the other relationships between A to B. This relationship implies the interaction between users in the graph database during the considered time.

All of the above relationships have been captured in a Neo4j instance.

5.2 Feature Analysis

In order to analyze and quantify the risk of the entities involved we observe the following exhaustive list of features as provided and inferred by the data that we have collected.

5.2.1 Tweet Features

Each of the tweets that have been collected provides a number of features. Table 5.1 summarizes these features, while we describe them in details below:

Table 5.1. Tweet Features list

Tweet Features	Description	Usage in HIV risk analysis
Tf-idf	Tf-idf of unigrams, bigrams and trigrams	Input for machine learning classifiers
urls	URLs in the tweet	Risk of the content in the url
hashtags	Hashtags in the tweet	Risk of known risky hashtags
user_mentions	User mentions in the tweet	Risk of known risky users
symbols	Symbols like cashtags in the tweet	Risk of known risky symbols
Images	Some tweets have images	Risk of the objects and themes of the images
possibly_sensitive	Flag provided by Twitter	Indicator of sensitive content in the URL attached
Location	Point cooridnates or bounding box of the location where the tweet was created	Helps in location based aggregations
created_at	Time the tweet was created	Helps in analysing the time of day for risky tweets
Related tweets	If the current tweet is a retweet, quotetweet or reply for another tweet	Risk of the related tweet
quote_count	Number of times the tweet got quoted	Popularity of the tweet
retweet_count	Number of times the tweet got retweeted	Popularity of the tweet
favorite_count	Number of times the tweet got favorited	Popularity of the tweet
reply_count	Number of times the tweet got replied	Popularity of the tweet

- **Tf-Idf** of unigrams, bigrams, trigrams, etc. in the tweet can be used as an input to a variety of machine learning models that can serve as tweet classifiers.
- **Known Risky Entities** like hashtags, URLs, user_mentions, and symbols that appear in tweets can be analyzed for the presence of HIV risk.
- **Images** in the tweet can be analyzed using Google Cloud Vision API for the presence of risky themes and objects.
- **Location** information in the Tweets, either in the form of a point coordinates or a bounding box can help track the riskiness corresponding to specific places in the territory.

- **Reply, Retweet, and Quotetweet** enable to link other tweets to the current one and can be used to calculate composite HIV risk from related tweets.

- **quote_count, retweet_count, reply_count, and favorite_count** help understand the popularity of a tweet and can be used to understand the potential spread of HIV risk.

- **possibly_sensitive** is a flag that is turned on when a link in the tweet has been identified to contain sensitive content. This is often a good indicator of HIV risk behavior.

- **created_at** indicates when the tweet was first created and can help calculating statistics about HIV risk at particular times of the day, or during major HIV at-risk events.

5.2.2 User Features

Similarly, each user that has been captured provides another set of user-level features. Table 5.2 summarizes them, and we describe them below:

Table 5.2. User Features list

User Features	Description	Usage in HIV risk analysis
user_timeline	Most recent tweets posted by the user	Ratio of risky to non-risky tweets
Location	Location of the user	Helps in location based aggregations
name/screen_name	User profile's names	Risk of the names
description	User's Description	Risk of the description
url	User's URL	Risk of the content in the URL
followers_count	Number of users who follow this user	Popularity of the User
friends_count	Number of users this user follows	Connections of the User
verified	Flag provided by Twitter	Indication of public interest in the profile
Profile image	Profile picture of the user	Risk of the objects and themes of the image
Background image	Background image of the profile	Risk of the objects and themes of the image

- **user_timeline** encodes the user activity and allows us to investigate the HIV risk of a person over time by looking at the ratio of risky to non-risky tweets.

- **Location:** we can use the location of the user in combination with the location of the users' tweets (see tweets features above) to determine where he/she are located. This could help analyze location-based HIV risk.
- **Names and Description** can be marked as risky in the case of the presence of risky themes.
- **Profile/Background Images** can be analyzed for risky themes with the help of the Google Cloud API.
- **verified** is a flag that tells if the user's account is of public interest. This helps to understand whether the user should be included in HIV risk analysis.
- **follower_count and friends_count** could help calculate the popularity of a user and analyze the spread of HIV risk.

5.2.3 Graph Relation Features

This set of features are derived from the connections across the captured users and tweets. These features are summarized in Table. 5.3 and explained below

- **Geographic Co-location and Temporal Co-occurrence** of tweets when they happen together indicate that the users are in the vicinity of each other around the same time. If tagged as risky users, this feature might indicate possible increased HIV risk.
- **Follows, Mentions, Retweets, and Quotetweets** indicate that a user shows interest in another one. If one of these relationships exist also in the reverse order, this implies a stronger connection, and it is likely that the users know each other in real life. If tagged as risky users, this feature might indicate possible increased HIV risk.
- **Conversation Chains** frequency and length indicate the strength of a transient connection between the users. If the tweets in these conversation chains belong to the risk buckets, we assume that this could be a risky conversation.

Table 5.3. Graph Features list

Graph Features	Description	Usage in risk analysis
Geographic Co-location	If 2 different users' risky tweets are located within one mile of each other	Indicates a user-user connection by location
Temporal Co-occurrence	If 2 different users' risky tweets are created within one hour of each other	Indicates a user-user connection by time
Follows relationships	User A → Follows → User B.	Indicates User A's interest in User B
	Does (User B → Follows → User A) ?	Indicates mutual interest
Mentions relationships	User A → Mentions → User B.	Indicates User A's interest in User B
	Does (User B → Mentions → User A) ?	Indicates mutual interest
	Number of times User A mentions User B	Strength of the interest
Conversation Chains	When a user's tweet acts as a reply for another user's tweet	Indicates a user-user connection
	Length of the conversation chains	Strength of the connection
	Number of time such conversations have occurred.	Strength of the connection
	Risk buckets of the tweets in the conversation	Type of risk the conversation carries
Retweet relationships	User A → Retweets → User B.	Indicates User A's interest in User B
	Does (User B → Retweets → User A) ?	Indicates mutual interest
	How many times does User A retweet User B ?	Strength of the interest
Quotetweet relationships	User A → Quotetweets → User B.	Indicates User A's interest in User B
	Does (User B → Quotetweets → User A) ?	Indicates mutual interest
	How many times does User A quotetweet User B ?	Strength of the interest

With the help of these rich sets of features, we hypothesize that multimodal risk scores can be calculated and used to drive prevention efforts. Current strategies to calculate similar scores are limited to a restricted number of variables. For instance, Hoenigl et al. created an online calculator for the HIV-risk of MSM that takes only four variables into consideration [33]: (1) the number of male partners, (2) the number of male partners with whom the person had CRAI (Condomless Receptive Anal Intercourse), (3) the number of male partners, that were HIV positive, with whom the person had CRAI, and (4) if the person suffered from any other STIs, like Syphilis, Gonorrhea, Chlamydia, etc.

We believe that the list of features that our data science pipeline can produce in real-time from Twitter would serve as a great improvement in the determination of such risk scores. In the next chapter, we will look into our attempt at procuring HIV risk scores based on Twitter data.

Chapter 6

Risk Scores

In order to better track and understand the risk of each tweet and user captured in our database we define the following HIV risk scores:

1. Risk Score for Tweets
2. Risk Scores for Users. We will discuss three types of user risk scores.
   (a) **Simple** User Risk Score: Risk of users based on their own Twitter activity
   (b) **Connection** User Risk Score: Risk of users based only on the risk of the people they interact/connect with on Twitter.
   (c) **Composite** User Risk Score: Risk of users based on their own Twitter activity and also based on the risk of the people they interact/connect with on Twitter.

While these scores need to be computed to understand risk on Twitter, the intervention will be driven by aggregate information at a higher level. As discussed in [7], care needs to be taken with calculating HIV risk, especially in terms of how these scores are communicated and utilized.

In the previous chapter, we looked into the available feature list at our disposal to develop a methodology to quantify the HIV risk of the users. In this chapter, we will hypothesize heuristic-based approaches to calculate these risk scores.

6.1 Supervised ML for Tweet classification

6.1.1 Labelled tweet data

We had conducted a Hackathon with the HIV focus group that helped us provide labels to about 600 tweets. We also have the dashboard as the source for labels. It gave us about 300 tweet labels. In total, we have 931 labeled tweets among which 128 were labeled *risky*.

6.1.2 Vectorization of Tweets

For this Supervised machine learning task, we need to create input features from a Tweet to the ML model to predict the output label for the Tweet. In order to make maximum use of the data available to us, we use the *bag-of-words* model and pre-trained *GloVe* [34] Word Vectors to create the input feature vector.

These are the following pre-processing steps taken to create a vector for a given Tweet:

1. Remove all punctuation in the Tweet.
2. Consider all the Hashtags as ordinary words (i.e. remove the # from the HashTags).
3. Convert all the *mentions* and other annotations to ordinary words by removing the @.
4. Gather labels for the images attached to the Tweet using Google's Cloud Vision API. Add these words too to the *bag of words* that represent the Tweet.
5. Use the resultant *bag of words* and the *Tf-Idf* [1] along with the GloVe pre-trained word vectors to create the final input vector based on a given Tweet.

$$X = \Sigma_{\forall word \in (BagOfWords(Tweet) \cap GloVe)} TfIdf(word) * w2v(word)$$

- $BagOfWords(Tweet)$ is the set of all the words representing the Tweet. This is procured as explained in steps 1-4 above.

[1] https://en.wikipedia.org/wiki/Tf–idf

- $TfIdf(word)$ is the *Tf-Idf* of the word as derived from the training data corpus with labelled Tweets.
- $w2v(word)$ is the 200-dimensional vector representation of the word as per the *GloVe* pre-trained word vectors, that were trained based on about 2 Billion Tweets worldwide.

Therefore, the vector X that has been generated in the final step as explained earlier is a 200-dimensional numerical vector that represents a Tweet. Each dimension in the vector is representative of an unnamed characteristic about the tweet. This 200-dimensional numerical vector serves as an input to our Supervised Machine Learning Classification model.

Advantages of using Pre-trained Word Vectors:

One of the very useful advantages of using pre-trained word vectors in our work is that we enhance the power of limited labeled data we have. Since the number of tweet labels is quite low (931 tweets), we do not have the vocabulary of all the terms that are HIV risky in the Twitter world. By using the GloVe word vectors pre-trained with about 2 billion tweets, we now have access to a broader vocabulary.

One of the disadvantages of directly using Tf-Idfs of the n-grams in tweets as features is that the feature vector would be very sparse. By using the pre-trained word vectors we are limiting the size of the tweets vectors to 200 and these vectors are not sparse.

6.1.3 Accuracy Measures:

Definitions

$$Precision^2 = \frac{tp}{tp+fp}$$

[2]https://en.wikipedia.org/wiki/Precision_and_recall

$$Recall^{3} = TruePositiveRate(TPR)^{4} = \frac{tp}{tp+fn}$$

$$FalsePositiveRate(FPR)^{5} = \frac{fp}{fp+tn}$$

The F1 Score[6] is the harmonic mean of *precision* and *recall*.

$$F_1 = 2.\frac{precision.recall}{precision+recall}$$

The ROC curve is created by plotting the *true positive rate (TPR)* against the *false positive rate (FPR)* at various threshold settings.

The Precision-Recall curve[7] is created by plotting the *Precision* against the *Recall* at various threshold settings

- tp : Number of *True Positives* as determined by the model
- tn : Number of *True Negatives* as determined by the model
- fp : Number of *False Positives* as determined by the model
- fn : Number of *False Negatives* as determined by the model

Precision vs Recall vs F1-Score

Since we expect the number of risk tweets to be very less compared to the total number of tweets, the original idea was to capture all the tweets that were risky by minimizing the *false negative* tweets and thus achieving a High **Recall**. But, we observed that while we were trying to maintain a high Recall we were compromising a lot on the **Precision**, and therefore we were ending up with an unexpected increase in the number of falsely labeled tweets as risky (i.e. *false*

[3]https://en.wikipedia.org/wiki/Precision_and_recall
[4]https://en.wikipedia.org/wiki/Sensitivity_and_specificity
[5]https://en.wikipedia.org/wiki/False_positive_rate
[6]https://en.wikipedia.org/wiki/F1_score
[7]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html

positive tweets). Therefore, in order to capture the best *Recall* while capturing the best possible *Precision*, we looked into capturing the best final *F1-Score* for the classification algorithm.

Receiver Operating Characteristic(ROC) Curves vs Precision-Recall(PR) Curves

In measuring the validity of a binary classification algorithm, the area under either of ROC or PR Curves can be used. In our case, we clearly have a class imbalance between the positive labels (*risky*) and the negative labels (*non-risky*). As stated in this journal [35] by Saito and Rehmsmeier the Precision-Recall plot is more appropriate while validating Binary Classifiers with imbalanced data, as compared to the *ROC Curves*. Therefore, we looked at the area under the Precision-Recall Curves during the validation process.

6.1.4 Classification Model Selection:

We have multiple machine learning based classification algorithms at our disposal. We will be looking into exploring the following algorithms for our purpose- Support Vector Machines (SVM) with *radial bias function(RBF)* kernel, and with *polynomial* kernel (with degree=3)[36], Logistic Regression with a *liblinear* solver [37], Random Forest classifier[38] and Adaboost Classifier[39]. We have used the python library *scikit-learn* [40] for the experimentation, application and validation of our machine learning models.

In order to choose the right model with the right hyper-parameters we follow the following steps:

1. Choosing regularization parameters for each individual algorithm using Precision-Recall curves

While using the SVMs or Logistic regression, the regularization parameter is the **C** parameter which is the penalty parameter for the error term. This means that the lower the parameter the stronger the regularization.

While using the Random Forest Classifier or the Adaboost Classifier, we determine the number of estimators (i.e. the parameter **n_estimators**) as part of this step. In the case of the

Table 6.1. Regularization parameters for each of the algorithm that we experimented with.

Algorithm	C-parameter	n_estimators
SVM with RBF kernel	9000	-
SVM with 3^{rd} degree polynomial kernel	20000	-
Logistic Regression (with a liblinear solver)	7	-
Random Forest Classifier	-	60
Adaboost Classifier (with Decision Tree Classifiers (max Depth = 5) as Base estimators)	-	120

Adaboost Classifier, the base estimators we used were instances of Decision Tree Classifier [41] with a maximum depth of 5.

We used the K-Fold validation technique (with K=20) to determine the regularization parameter that gave the best *Area under the Precision-Recall curve* based on *test* data for each of the algorithm types. Fig. 6.1 shows the plots of change in Area under the Precision-Recall curves with the changes in the hyper-parameter in both the Training and Testing data for each of the algorithms.

The regularization parameters determined as part of this process are detailed in Table 6.1

Fig. 6.2 shows the Precision-Recall Curves for each of the individual algorithms with the best regularization parameters.

2. Selection of the best model and its optimal probability threshold based on F1-Scores

For each of the individual models with the optimal hyper-parameters (as captured in the previous step), we conducted K-Fold cross-validation (with K=25) in order to capture the optimal threshold that gave the highest F1-score as shown in table 6.2

Fig. 6.3 shows the change in the F1-scores with the change in the probability thresholds for each of the individual algorithms with their best determined hyper-parameter.

Since the 3*rd*-degree polynomial based SVM with the hyper-parameter $C = 20000$ and threshold 0.1958 gave the best F1-score of 0.3517, we trained it with the full training data and used it to classify the tweets in our Experimental setup.

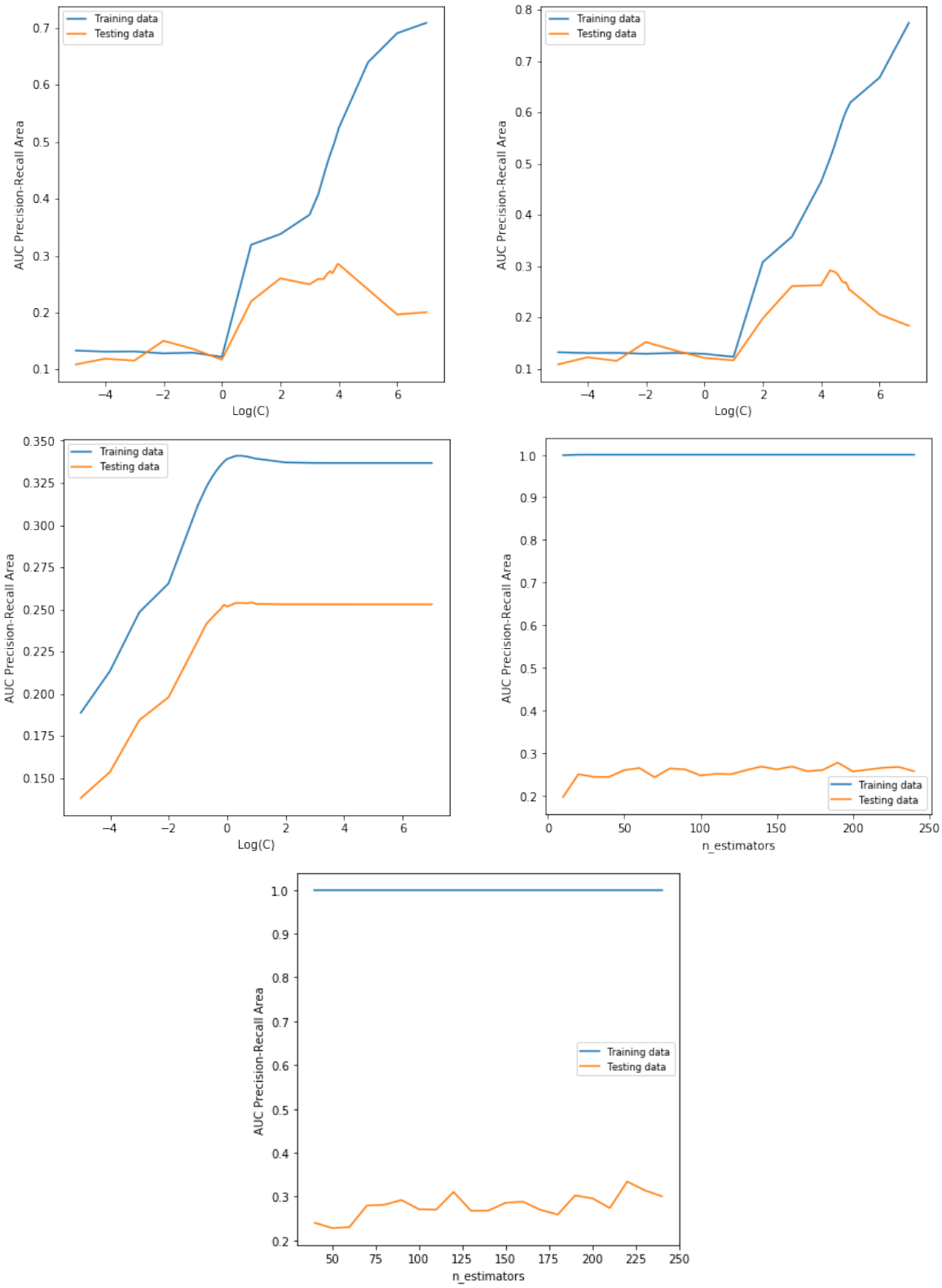

Figure 6.1. Change in AUC for the Precision-Recall curve in each of the algorithms with the change in their Hyper-parameter. *Top-left*:SVM with RBF kernel; *Top-right*: SVM with 3^{rd} degree polynomial kernel; *Mid-left*: Logistic Regression; *Mid-right*: Random Forest Classifier; *Bottom*: Adaboost Classifier

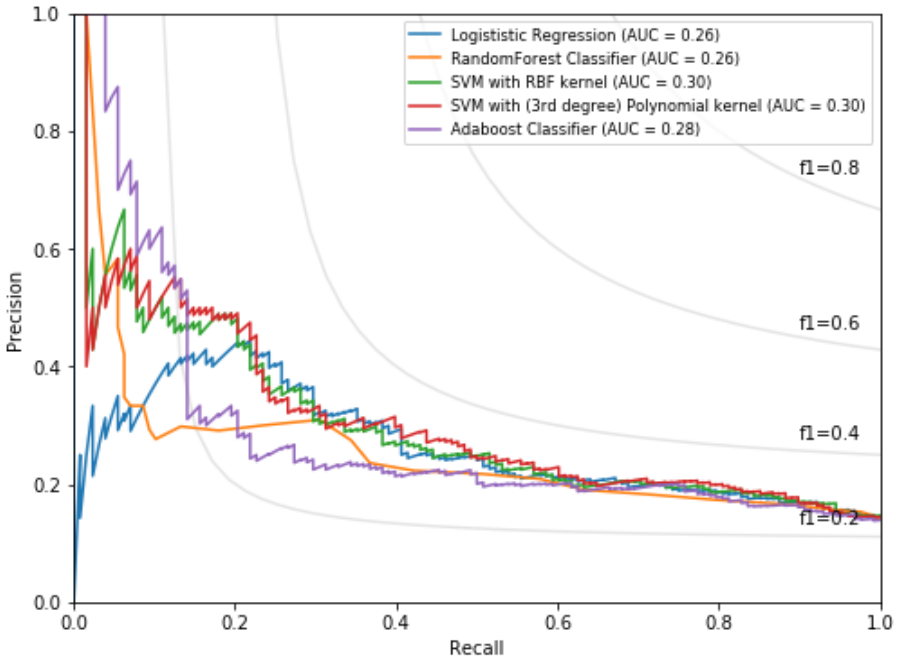

Figure 6.2. Precision Recall curves for each of the models

Table 6.2. Threshold and F1-scores of the models with best regularization hyper-parameter for each Algorithm

Model with best regularization param	**Probability Thresholds**	**F1-score**
SVM with RBF kernel	0.1736	0.3393
SVM with 3^{rd} degree polynomial kernel	0.1958	**0.3517**
Logistic Regression (with a liblinear solver)	0.6269	0.3415
Random Forest Classifier	0.2333	0.3071
Adaboost Classifier (with Decision Tree Classifiers (max Depth = 5) as Base estimators)	0.0015	0.3168

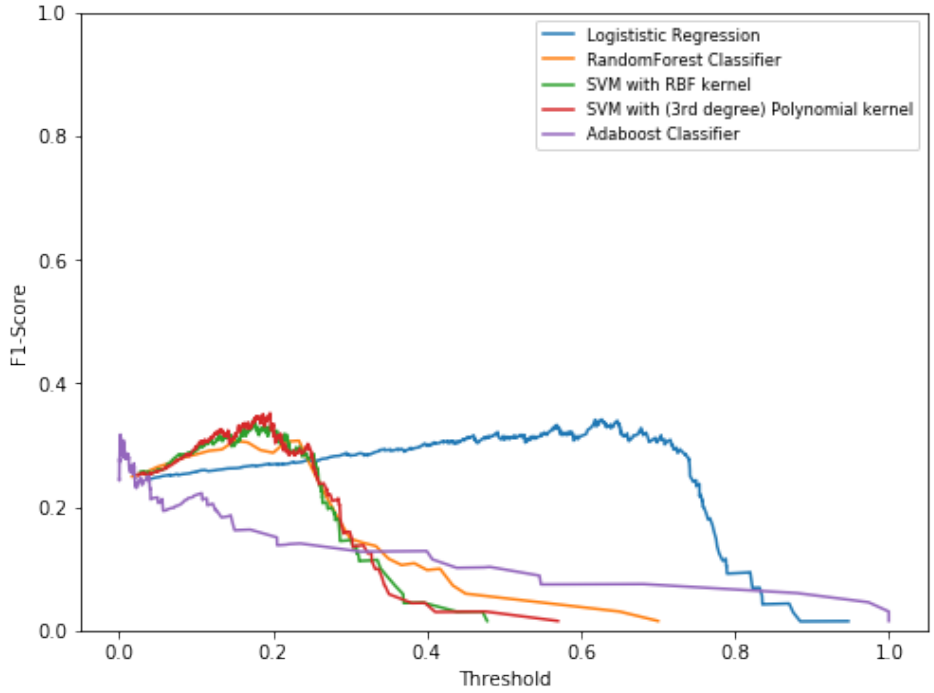

Figure 6.3. Change in F1 scores with change in probability Threshold for each of the models

The risk probability calculated from the machine learning model for each of the tweets is the **risk score** of a Tweet.

6.2 User Risk Scores

All the user risk scores are time-dependent, which means that they change with the passage of time because the behavior of a person keeps changing from one time period to another. Note that, we do not have a ground truth about the risk of the users in our experimental data. Therefore, we hypothesize the following ideas in the hope to be validated in the future.

6.2.1 Simple risk score

This risk score for a user depends on his/her own activity on Twitter. We look solely at the tweets of a person in the given time period and determine his/her risk. In our case, we take the ratio of risky tweets to the total number of tweets tweeted by a person in the given time

period. This ratio is always less than or equal to 1.

$$RiskScore_{u\_simple} = (u\_RiskTweetCount)/(u\_TotalTweetCount)$$

- $u\_RiskTweetCount$: This is the number of tweets by user u in the given time-period that were predicted to be risky by the most optimal ML model described in the previous section.
- $u\_TotalTweetCount$: This is the total number of tweets by user u in the given time-period.

Note that, this risk score is undefined if the person doesn't have any tweets in that time period. Fig. 6.4 shows the distribution of the simple risk score in our experimental data.

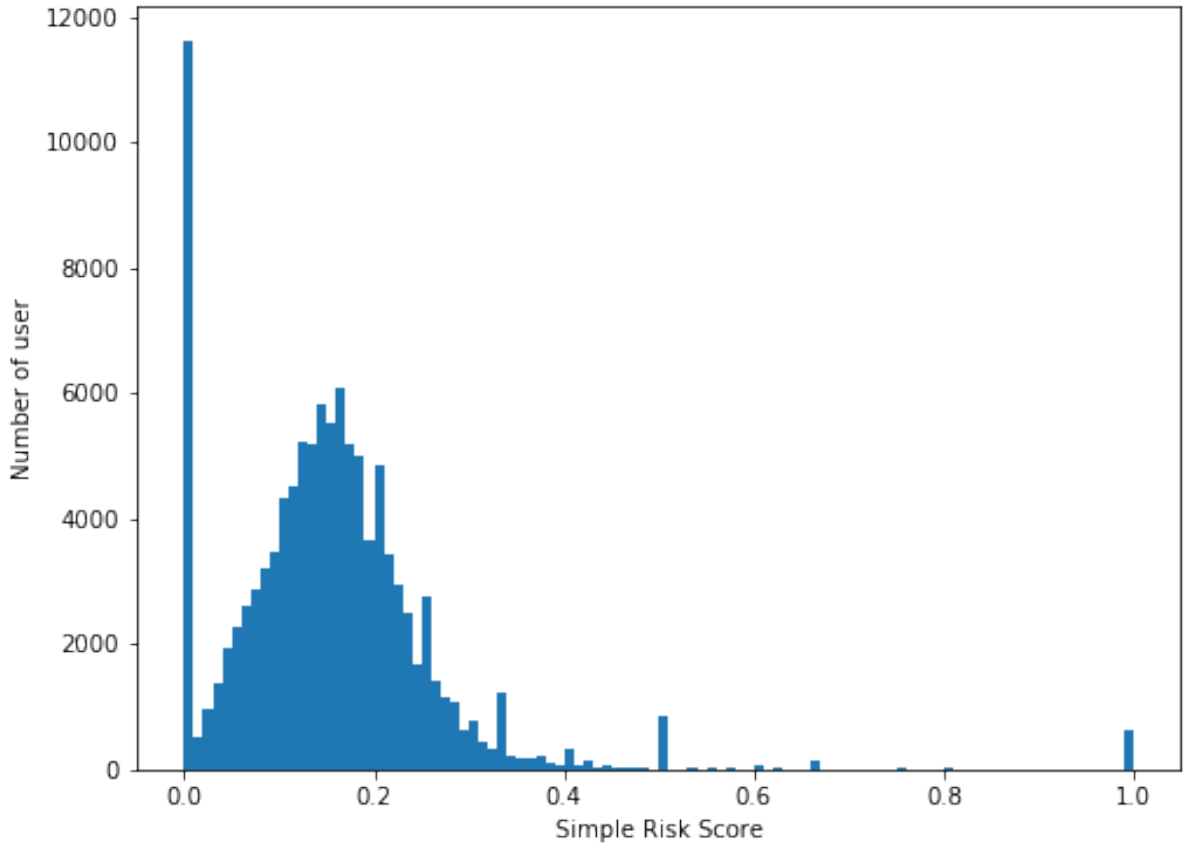

Figure 6.4. This is the distribution of simple user risk scores calculated for the users (count = 110,189) with at least one tweet in the experimental data

6.2.2 Connection risk score

Since HIV is contracted from one individual to another, it is important to account for the risk of the people connected with a person while calculating his/her risk.

$RiskScore_{u\_connections}$ is the HIV risk of user u based on his/her twitter connections.

$$RiskScore_{u\_connections} = \frac{\Sigma_{\forall n \in Neighbors(u), n \neq u} ConnectionStrength(u,n).RiskScore_{n\_simple}}{\Sigma_{\forall n \in Neighbors(u), n \neq u} ConnectionStrength(u,n)}$$

- The set $Neighbors(u)$ is the set of all the nodes to which there is an edge from the node u. Note that, there could be cases where $u \in Neighbors(u)$, therefore in the above expression we only look at the *neighbors* that are not the same node as u.

- The $ConnectionStrength(u,n)$ explains how well connected user u is with user n. In our case, we use the *count* parameter of the $ACTIONED$ relationship as the connection strength. Note that, the *count* parameter is time-dependent because the $ACTIONED$ edges in the graph are created based on the number of times the tweets of User A interacted with User B during a given time period.

- $RiskScore_{n\_simple}$ is the simple risk score of u's neighbor n in the connections graph.

Note that, like the *Simple Risk Score* the *Connection Risk Score* is also always less than or equal to 1.

Fig.6.5 shows the distribution of the connection risk score in our experimental data.

Fig.6.6 shows the scatter-plot between the connection and simple risk scores. We observed a positive correlation of nearly 0.3 between the two scores. This is a significant observation because it shows that users' own behavior on Twitter shows a weak positive correlation to the behavior of the people they are connected to on Twitter.

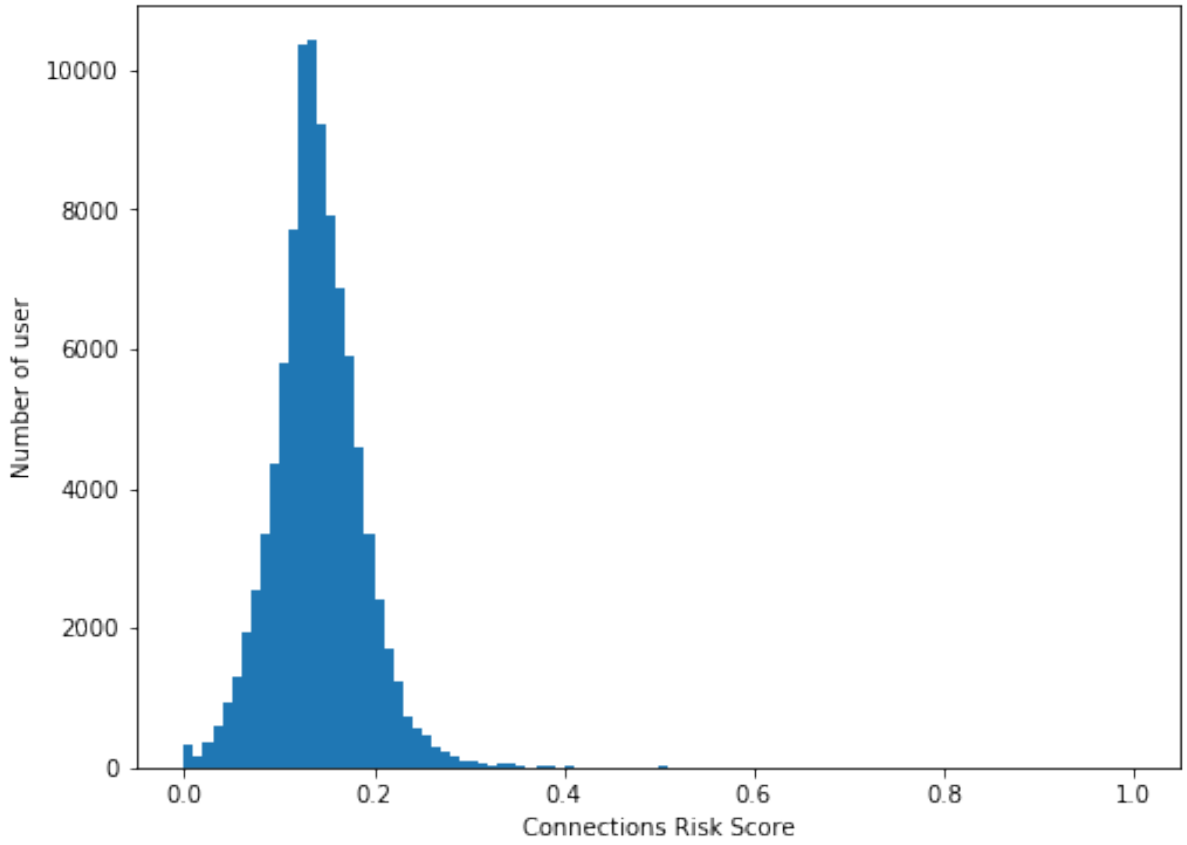

Figure 6.5. The distribution of connections based user risk scores calculated for the users (count = 96742) with at least one *ACTIONED* relationship originating from them

6.2.3 Composite risk score

The composite risk score of a user depends on the user's own activity on Twitter as well as the risk of the connected users that the user has interacted with during the specific time period. The connections in our case are defined by the *ACTIONED* relationship as explained earlier in Chapter 5.

The Composite risk score of a user $RiskScore_{u\_composite}$ is given as:

$$RiskScore_{u\_composite} = \alpha.RiskScore_{u\_simple} + (1-\alpha).RiskScore_{u\_connections}$$

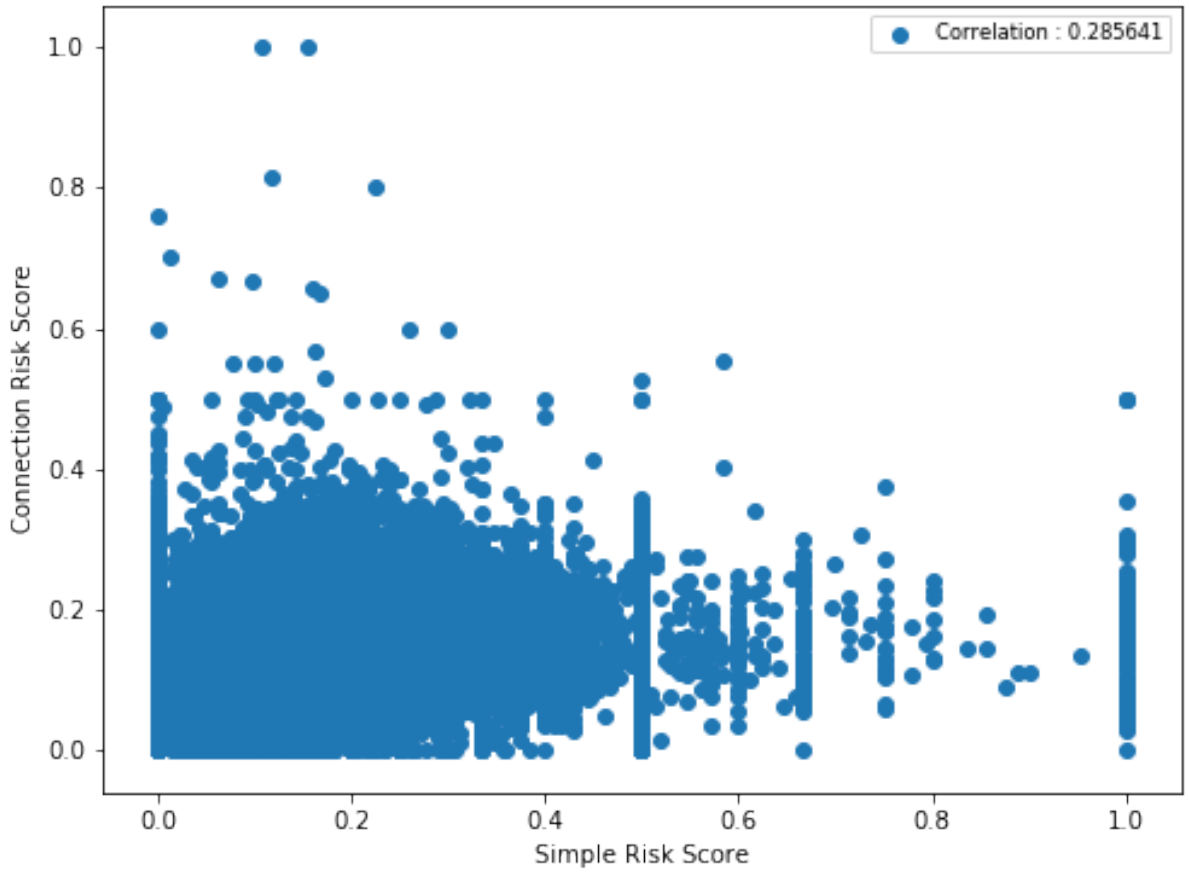

Figure 6.6. A Scatter plot showing the correlation between connection and simple risk scores.

- $RiskScore_{u\_simple}$ is the simple risk score of the user
- $RiskScore_{u\_connections}$ is the connection risk score of the user.
- α is value such that $0 <= \alpha <= 1$. This parameter could be learned or can be set manually.

Since both the simple and connection risk scores are between 0 and 1 (inclusive), their weighted average is also between 0 and 1 (inclusive). Therefore all the user risk scores are between 0 and 1.

6.2.4 Identifying Risky networks

One of the many advantages of this approach is that we can identify live risky networks in the user graph. The network will be formed by users who have more than a threshold of risk score τ_r and *ACTIONED* links between them with a count more than a threshold of τ_c. Note that, we can restrict the bond to be stronger amongst users by looking at only the users who have *ACTIONED* relationships in both directions

Cypher Query :

$$MATCH\ \big(u1:USER\big)-\big[r1:ACTIONED\big]->\big(u2:USER\big),$$
$$\big(u1:USER\big)<-\big[r2:ACTIONED\big]-\big(u2:USER\big)\ where$$
$$u1.id <> u2.id\ and\ u1.simple\_risk\_score >= 0.25\ and$$
$$u2.simple\_risk\_score >= 0.25\ and\ r1.count >= 5\ and\ r2.count >= 5\ return *$$

Note that in this Cypher query, $\tau_r = 0.25$ and $\tau_c = 5$. Fig. 6.7 shows the results of the Cypher query above. Notice that, from the results, we can extract major networks of users that actively participate in HIV risk behavior as per their Twitter data. This can help the Public health departments to help fabricate customized measures focused on creating preventive awareness in those individual networks.

Acknowledgements

We would like to express my gratitude to Prof. Lawrence Saul who helped us in developing and validating the ideas and approaches that are discussed in this chapter.

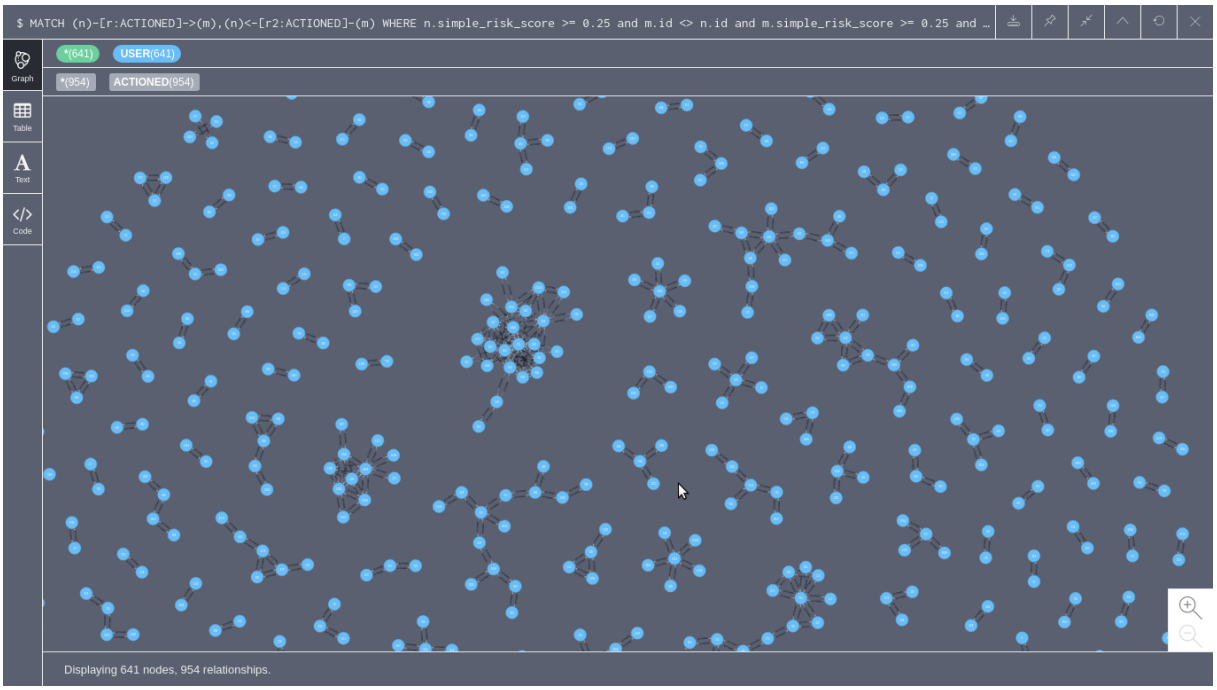

Figure 6.7. $u1 \Rightarrow ACTIONED \Rightarrow u2$, $u2 \Rightarrow ACTIONED \Rightarrow u1$ (i.e. bi-directional $ACTIONED$) relationships extracted from the Neo4j instance setup for the experimental data using the above Cypher Query.

Chapter 7

Conclusion and Future work

In this thesis, we presented a flexible Data Science platform and architecture and showed how we can extract HIV risk behavior from Twitter. This large number of multi-modal Twitter features informs the definition of a new kind of HIV risk score that goes beyond current approaches.

The modular infrastructure presented here is the starting point for a myriad of research endeavors and can easily be extended in many directions. With the help of the near real-time capture of HIV risk trends in the Twitter population, this is a great step ahead in the field of digital epidemiology. The public health department will be the ultimate consumer of the information that is being generated by this pipeline.

Although we implemented a major part of the pipeline and established a proof of concept for the ideas discussed earlier, there still is a long way ahead of us for getting fruitful results that are helpful for the society. Here are some research areas that this project could take up in the future.

7.1 Collection of more reliable labelled data:

The effectiveness of the pipeline heavily depends on how robust the supervised machine learning models are. That, in turn, depends on the amount of reliable labeled data we can use to train the models.

7.1.1 Tweet Labels:

With the current number of labeled data, it is clear that we cannot encompass the large variety of tweets that we see on a daily basis and thus we end up with inaccurate risk scores and labels for the tweets. This inaccuracy makes its way to the calculation of user risk scores as well. This issue can be avoided by the collection of more tweet labels.

Using n-grams for tweet vectors:

Although the pre-trained *GloVe* word vectors help us with encountering the problem of limited training data, we use the *bag-of-words* model for unigram, and only the terms that are present in the pre-trained model can be translated to vectors. If we have more labeled data, we can experiment with n-grams (instead of unigrams), along with their *Tf-Idf* values while training the classification models

Tweet image labels:

Currently, we are converting images into text with the help of the Google's Cloud Vision API as explained in Chapters 3 and 6. But in the presence of a large number of labels for the tweet images we see in our own tweet corpus, we can create our own image classifier whose output can be used as an input for the tweet classification model.

7.1.2 User Labels:

Even though, we have hypothesized the User risk scores in Chapter 6 we have not validated the results of the hypothesis. This is only possible if we know the actual risk scores of some of the users in our database. If we had a list of known *risky/non-risky* users we can use supervised machine learning based regression models to estimate the risk of a user from his/her Twitter activity and connections.

While collecting data for user profiles, if the reasons for risk are captured for a risky user, it would help in validating the models for user risk scores that are presented in this thesis.

7.1.3 Data collection methods:

We have currently adopted the use of the crowd-sourcing dashboard and domain-experts for the collection of labeled data. The user information captured in the crowd-sourcing dashboard can serve in the validation of the user risk score models as discussed earlier. The crowd-sourcing dashboard could also capture labels for Tweet images in the future. This could help in the design of a new set of image classifiers that could help contribute to the risk model for tweets.

In the future, we can adopt the use of Amazon Mechanical Turk [1] to gather user based and tweet based labels, while capturing the reasons for risk to enhance our labels and risk models.

7.1.4 Individual models and networks for each risk category:

In the presence of thousands of labeled data-points for users and tweets, we can design supervised machine learning models for each risk bucket in Chapter 3 individually and form individual risk user networks for each of the risk categories. This could help with better insights into the population that is involved in each of the risk behaviors, individually. For example, a Data pipeline designed for tracking *Drug Use* could be used solely for tackling the drug problem in the local community. With independent data pipeline for each risk behavior, the public health departments can focus on more and better tailored preventive measures for fighting HIV.

7.2 Data Visualization

The main objective of this project is to gather actionable inference about the changes in the HIV risk among the Twitter population. The face of these efforts should be a Data visualization module that could be used by the public health departments and the broader research community. These visualization efforts should take immense care to preserve the privacy of individual users while aggregating the data for a broader population. This could involve visualizations based on Region, Race, Gender and other demographic parameters. Work needs to be done to create an

[1] http://www.mturk.com/

easily navigable dashboard that displays the statistics about the huge number of users and tweets in the databases while protecting the privacy and anonymity of the individual users.

7.3 Exploring other data sources

Though the current data collection and storage pipeline are tailored towards the content that is being created in Twitter, we believe once we gather enough confidence on our hypotheses by gathering more data and results, we can look forward to expanding these efforts on other data sources like *Instagram, Facebook, etc*. We would also look at Gay dating websites such as *Grindr* as future data sources to our project.

Although the current approach is not devoid of issues, we believe that after implementing the proposed ideas this work will spearhead the research in HIV-risk digital epidemiology and allow for better data-driven prevention interventions. Our platform will also serve as a foundation for future research aimed at capturing real-time trends in social media.

When the aggregated anonymous information is available to the public health department, it can serve as a basis for their efforts to fight the spread of HIV by educating the citizens in the most appropriate ways and preventing the risk in a timely manner. This is a huge step forward compared to the current reactive approaches that are based on information available in Census data and in functioning HIV clinics and testing centers. With the availability of HIV risk data for the local Twitter population, the administration can keep track of the trends in HIV risk while understanding the effectiveness of the measures that are being taken to curb the risk on a macro-scale.

Bibliography

[1] UNAIDS, "Global HIV & AIDS statistics – 2018 fact sheet," *Ending the AIDS Epidemic*, July 2018. http://unaids.org/en/resources/fact-sheet.

[2] CDC, US Center for Diseases Control, "HIV Surveillance Report: Diagnoses of HIV Infection in the United States and Dependent Area," *HIV Surveillance Report 2016*, vol. 28, 2016. https://www.cdc.gov/hiv/statistics/overview.

[3] H. G. Setty, M. Kumar, and I. K. Hewlett, "Point of care technologies for hiv," *AIDS research and treatment*, vol. 2014, 2014.

[4] M. Dredze, "How social media will change public health," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 81–84, 2012.

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

[6] N. Thangarajan, N. Green, A. Gupta, S. Little, and N. Weibel, "Analyzing social media to characterize local hiv at-risk populations," in *Proc. Wireless Health 2015*, pp. 11–20.

[7] N. Weibel, P. Desai, L. Saul, A. Gupta, and S. Little, "HIV Risk on Twitter: the Ethical Dimension of Social Media Evidence-based Prevention for Vulnerable Populations," in *Proc. HICSS 2017*.

[8] J. W. Ayers, T. L. Caputi, C. Nebeker, and M. Dredze, "Don't quote me: reverse identification of research participants in social media studies," *Nature Digital Medicine*, vol. 1, no. 1, p. 30, 2018.

[9] G. Linden, B. Smith, and J. York, "Amazon recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, no. 1, pp. 76–80, 2003.

[10] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proc. RecSys 2016*, 2016.

[11] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, and V. Ramavajjala, "Smart reply: Automated response suggestion for email," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 955–964, ACM, 2016.

[12] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. K-CAP*, pp. 70–77, 2003.

[13] M. Salathé, "Digital epidemiology: what is it, and where is it going?," *Life sciences, society and policy*, vol. 14, no. 1, p. 1, 2018.

[14] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, 2013.

[15] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on twitter," *Journal of Research in Personality*, vol. 46, no. 6, pp. 710–718, 2012.

[16] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Proc. PASSAT/SocialCom 2011*.

[17] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proc. SIGKDD 2011*, pp. 430–438, 2011.

[18] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[19] N. B. Silva, R. Tsang, G. D. Cavalcanti, and J. Tsang, "A graph-based friend recommendation system using genetic algorithm," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pp. 1–7, 2010.

[20] H. J. Hospers, G. Kok, P. Harterink, and O. de Zwart, "A new meeting place: chatting on the internet, e-dating and sexual risk behaviour among dutch men who have sex with men," *Aids*, vol. 19, no. 10, pp. 1097–1101, 2005.

[21] N.-Y. Ko, S. Koe, H.-C. Lee, C.-F. Yen, W.-C. Ko, and S.-T. Hsu, "Online sex-seeking, substance use, and risky behaviors in taiwan: results from the 2010 asia internet msm sex survey," *Archives of sexual behavior*, vol. 41, no. 5, pp. 1273–1282, 2012.

[22] M. McFarlane, S. S. Bull, and C. A. Rietmeijer, "The internet as a newly emerging risk environment for sexually transmitted diseases," *Jama*, vol. 284, no. 4, pp. 443–446, 2000.

[23] S. S. Bull, D. K. Levine, S. R. Black, S. J. Schmiege, and J. Santelli, "Social media–delivered sexual health intervention: a cluster randomized controlled trial," *Am. J. Prev. Med.*, vol. 43, no. 5, pp. 467–474, 2012.

[24] N.-Y. Ko, C.-H. Hsieh, M.-C. Wang, C. Lee, C.-L. Chen, A.-C. Chung, and S.-T. Hsu, "Effects of internet popular opinion leaders (ipol) among internet-using men who have sex with men," *Journal of medical Internet research*, vol. 15, no. 2, p. e40, 2013.

[25] M. E. Ireland, Q. Chen, H. A. Schwartz, L. H. Ungar, and D. Albarracin, "Action tweets linked to reduced county-level HIV prevalence in the United States: Online messages and structural determinants," *AIDS and Behavior*, vol. 20, no. 6, pp. 1256–1264, 2016.

[26] S. D. Young, G. Szekeres, and T. Coates, "The relationship between online social networking and sexual risk behaviors among men who have sex with men (msm)," *PloS one*, vol. 8, no. 5, p. e62271, 2013.

[27] C. Buntain and J. Golbeck, "This is your twitter on drugs: Any questions?," in *Proc. WWW 2015*, pp. 777–782, 2015.

[28] X. Yang and J. Luo, "Tracking illicit drug dealing and abuse on instagram using multimodal analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 4, p. 58, 2017.

[29] P. J. Desai, *PIRCNET: A Data Driven Approach to HIV Risk Analysis*. PhD thesis, UC San Diego, 2016.

[30] A. Mohan, *PIRC-Net: Twitter-based on demand public health framework for HIV risk estimation*. PhD thesis, UC San Diego, 2017.

[31] Joshua Roesslein, "Tweepy Documentation, Revision 7b10fc45," 2018.

[32] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proc. EMNLP 2008*, pp. 254–263, 2008.

[33] M. Hoenigl, N. Weibel, S. R. Mehta, C. M. Anderson, J. Jenks, N. Green, S. Gianella, D. M. Smith, and S. J. Little, "Development and validation of the san diego early test score to predict acute and early hiv infection risk in men who have sex with men," *Clinical Infectious Diseases*, vol. 61, no. 3, pp. 468–475, 2015.

[34] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[35] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[38] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] L. Breiman, *Classification and regression trees*. Routledge, 2017.