Classifying HIV risk tweets using tweets from San Diego county

Narendran Thangarajan
University of California, San Diego
La Jolla
California, USA.
naren@ucsd.edu

ABSTRACT

This project aims to characterize and classify tweets that show users exposing HIV risk behaviour through their tweets on the social networking site Twitter. A labeled dataset obtained from doctors in UCSD's Anti Viral Research Center (AVRC) was used as the dataset. To get a better understanding of the data collected and to build a good classification model, a series of exploratory data analysis (EDA) experiments were performed on the training dataset. The EDA phase of the project revealed information on relevant and irrelevant features as expected. Then a comparative study is performed on classification using models built using logistic regression and Support Vector Machines. We find that logistic regression trumped over Support Vector Machines when the the domain specific terms collected from domain experts were made part of the feature set.

Categories and Subject Descriptors

[**Data Analysis**]: Exploratory Analysis; [**Machine Learning**]: Logistic Regression, Support Vector Machines

General Terms

Data Analysis, Machine Learning

Keywords

Twitter, HIV, HIV-risk tweets

1. INTRODUCTION

The advent of the internet in early 1990's allowed people to publish content online leading to a "Read" web. Then gradually the concept of blogs and collaboration was introduced allowing the internet to become a huge warehouse of people's opinions and ideas leading to a "Read-Write" web. The "Read-Write" web introduced the idea of social networks. Right now, social networks like Facebook, Twitter and LinkedIn are considered a mainstream platform for connecting with people across the globe. This brings in challenges in terms of scalability since each of these websites has to handle a deluge of requests hitting their loadbalancers every second. However, this also means that there is an untapped gold mine of people's opinions which when analyzed can give a great deal of insight to improve health care, advertising, natural disaster relief efforts etc. Along the same lines, the goal of this project is to tap the public tweets posted by twitter users to characterize and classify tweets that exhibit HIV risk behaviours.

2. DATA COLLECTION

There are are around 210 notable social networking sites starting from 43 things, a social network for goal setting and achievement to Zooppa, an online community for creative talent. In this project, a social network analysis is performed on Twitter mainly because of the earlier results published by Dr. Sean et. al. in their paper [8]. They had validated their approach by cross-checking the results obtained from Twitter to the real-world HIV spread information from AIDSVu organization.

Well known social networking sites like Twitter provide APIs to programmatically access their data instead of scraping their websites. TwitterâĂŹs engineering team provides Streaming APIs that provide third-party developers low latency access to TwitterâĂŹs global stream of tweets data. Twitter provides three kinds of Streaming APIs.

1. Sample hose
2. Fire hose
3. Filter hose

Streaming APIs create a long standing connection between the client and the server and stream the incoming tweets to the clients that have subscribed to those tweets. The sample hose provides tweets from all over the world at the rate of 70 tweets per second. Since this project was focused on San Diego county, the filter hose was used to get the geo-tagged tweets emitted from the bounding box across the SD county alone. This collects around 40 tweets per minute using this filter API. As this report is being written, 3,400,000 tweets have been collected so far just from SD county. The code for data collection was written prior to this project. It was slightly tweaked to cater to this project's requirements.

2.1 Data Collection Architecture

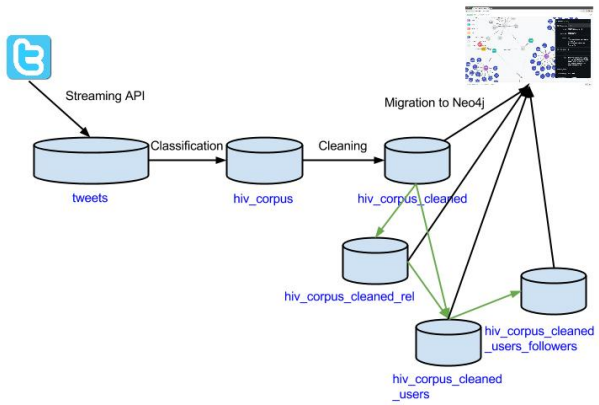

Figure 1: Data Collection Architecture

The tweets collected using the Streaming API are pushed on to a mongoDB collection named tweets. The reason for choosing mongoDB as the data store is mainly due to the rate at which we anticipated the tweets to be streamed. Next, from tweets collection, a smaller corpus of tweets is created by classifying the tweets based on the presence of certain HIV risk words. These HIV risk words were prepared with the help of Dr. Susan Little and Dr. Nella Green from UCSDâĂŹs Anti Viral Research Center. The risk words fall into 5 major categories.

1. Drugs
2. Sex
3. Sex Venues
4. Homosexual Terms
5. Sexually Transmitted Infections

3. DATA CLEANING

Most of these HIV risk words were derived from urban dictionary and apply as risk words only in certain contexts. Since a simple text match was used on these risk words to filter tweets, a lot of noise was found in the initial tweet collection. The observation was that for every risk word, we could find a set of words that can co-occur with them and be indicative of whether the tweet exhibits HIV risk behaviour or not. Such co-occurring words can be used in two ways: a)To filter out a certain tweet which might be noisy b) To filter in only matching tweets and filter out all other tweets for that particular risk word. We call words belonging to the prior as exception words and those to the latter as inclusion words.

The data cleaning phase led to reduction of the tweets by a significant 60%. Since the exception lists and the inclusion lists are being revamped on a regular basis as and when we learn more about the domain, the cleaning process is run as a batch process every 3 hours.

4. EXPLORATORY DATA ANALYSIS

Next, the following statistics about the social network graph were derived to get a better understanding of the Twitter sub-network derived from tweets that show HIV risk behaviour. We will call those tweets as HIV-risk tweets from now on. To ensure if the HIV-risk tweets alone show a different characteristic, its required to perform the EDA on both the HIV risk tweets and the whole Twitter dataset. The following charts show how the 11205 tweets are distributed based on different criteria.

4.1 Distribution of tweets

4.1.1 By time of day

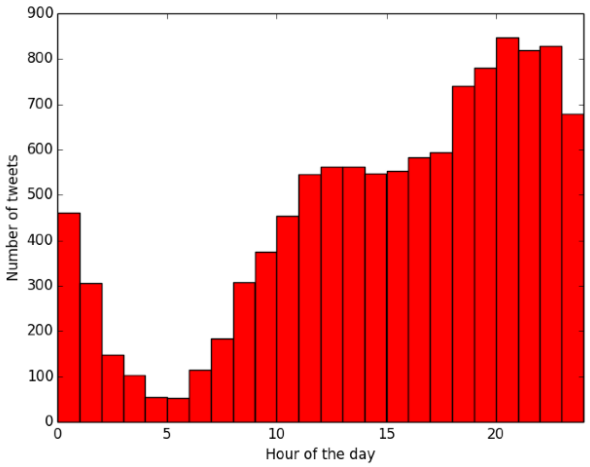

Figure 2: HIV risk tweet distribution with time of day

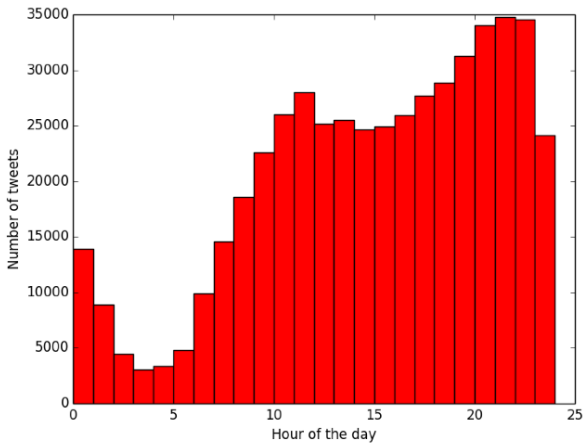

Figure 3: Tweet distribution with time of day

Firstly, as shown in the figure above, it looked like the HIV risk tweets were more pronounced towards the evening times. It hits a peak of 848 tweets from 8 pm to 9pm everyday and hits a low of 53 tweets on average from 5 am to 6 am. However, performing the same analysis on the set of all

tweets revealed that that pattern is because the underlying tweet distribution by itself has that characteristic.

So the distribution in the full dataset doesn't reveal a potential feature for our classification model. The next guess was to see if there is a pattern in the day of the week with the HIV risk tweeting behaviour.

4.1.2 *By day of the week*

As we observed with the time of day, the HIV risk tweet distribution and the actual underlying tweet distribution change in the same way. This can be see in the following figures.

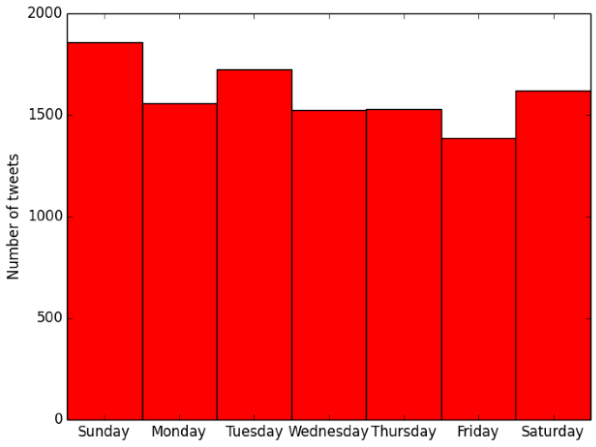

Figure 4: HIV risk tweet distribution with day of week

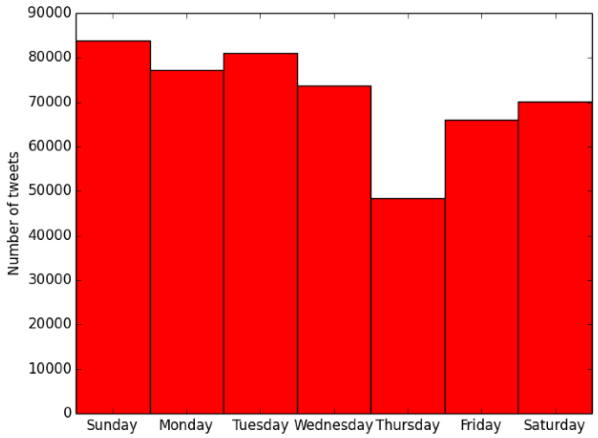

Figure 5: Tweet distribution with day of week

4.2 Degree distribution of users

4.2.1 *Based on number of tweets tweeted by the user*

The next aspect to explore was to check the total number of users who are actively involved in tweeting HIV risk tweets. If this is a small fraction, then it would be interesting to see the communities and cliques surrounding these specific users. As guessed, the data showed that there are only very few active users.

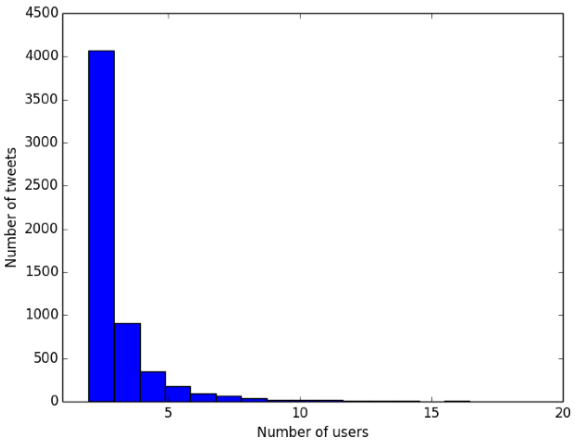

Figure 6: Degree distribution of HIV-risk tweets across users

Once again, when this was cross-verified with the underlying tweet distribution, it doesn't reveal anything interesting.

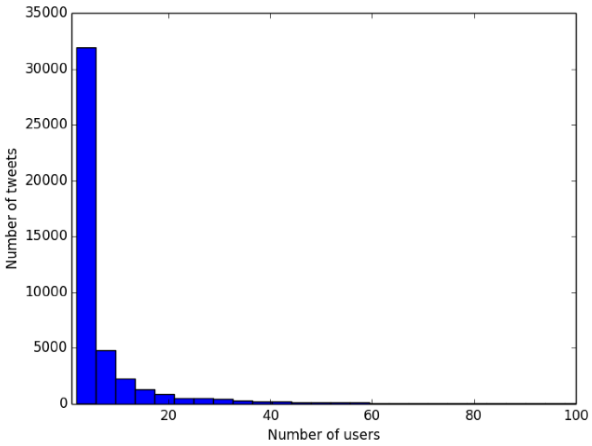

Figure 7: Degree distribution of tweets across users

4.2.2 *Based on number of tweets mentioning a user*

In Twitter, there are three major indicators of a twitter user's influence

1. Number of followers
2. Number of tweets mentioning that user.
3. Verified status

For instance, celebrities generally have a "verified" status, have a lot of followers and each tweet they post results in a deluge of responses from followers which eventually translate to "mentions" in twitter. There are organizations like

AIDSvu and the UCSD AVRC center on Twitter that post tweets which could contain the same risk words used by an actual HIV risk user, however, the context is totally different. To identify such cases, the feature set should include features that identify the influence of the twitter user who posted the tweet being classified. The following chart shows the power-law distribution found in the number of mentions in HIV risk tweets. This shows that the concept of influence remains the same as in the main Twitter social network.

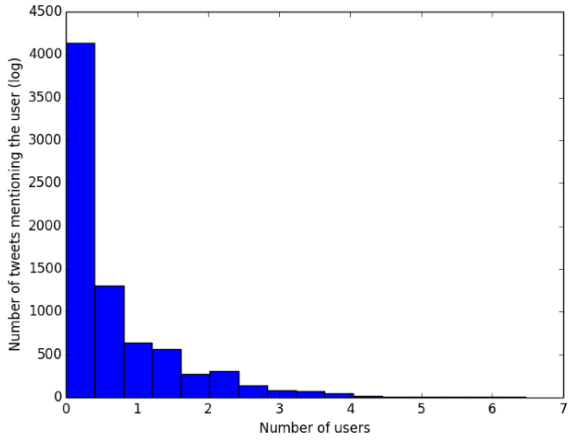

Figure 8: Degree distribution of mentions in HIV-risk tweets across users

4.2.3 Based on the length of the tweet

The next question was if the length of the tweet told anything interesting about the HIV-risk aspect of the tweet. Following are the distributions with respect to HIV-risk tweets and then with all the tweets collected.

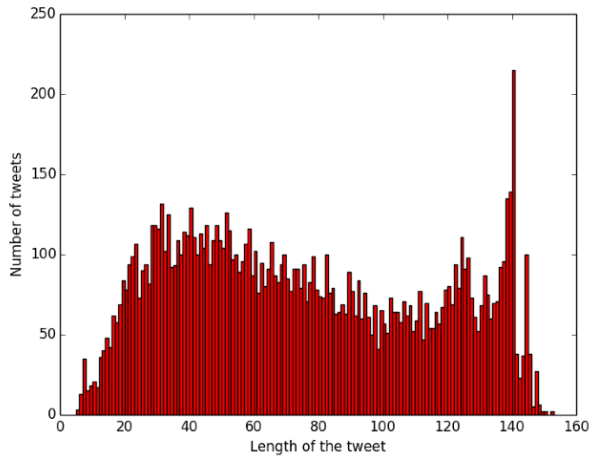

Figure 9: Degree distribution of HIV-risk tweets based on tweet length

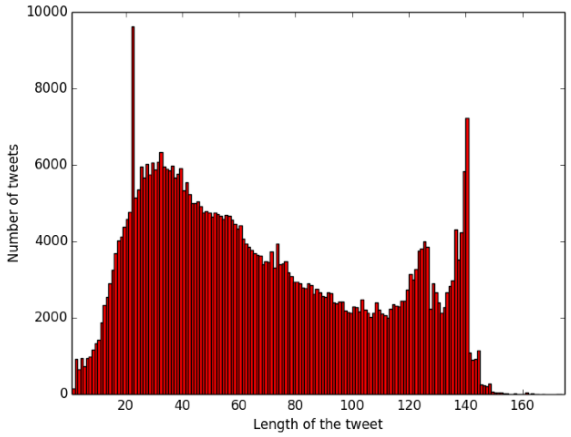

Figure 10: Degree distribution of tweets based on tweet length

There is not any marked difference between the tweet distributions. A couple of interesting things to note here.

1. There is a sudden spike for tweet length 22 in the global tweets dataset. On further exploration, this seems to be caused by tweets generated by customers' checkins into clubs in Hillcrest. So they can be safely avoided since we have already factored the presence of Sex Venues names in the tweet.

2. There are tweets which are longer than 140 characters. This is interesting because tweets are supposed to be limited to 140 characters. On further investigation, it became obvious that the emoticons are considered as a single character according to twitter, however, when we ask for string length, each emoticon contributes to two counts.

4.3 Distribution of tweets across the different HIV risk buckets

It was interesting to find the major fraction of HIV risk tweets just falling within Homosexual terms bucket and the Drug Bucket. This might be a good indicator of why AIDS is most predominant among Men who have Sex with Man (MSM) in San Diego county.

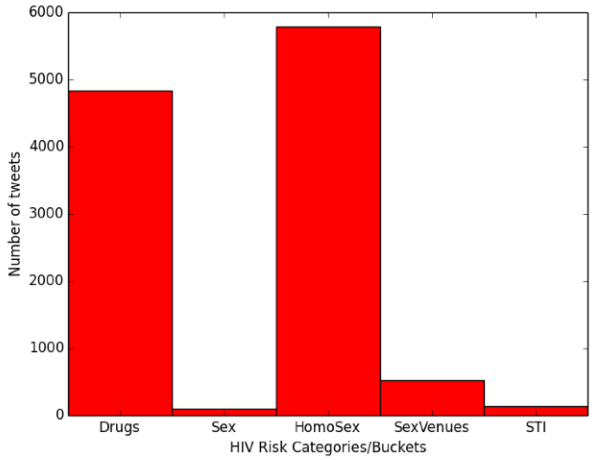

Figure 11: Degree distribution of tweets based on risk buckets

This gives the idea that the presence/absence of certain drug related or homosexual relationship terminologies could provide a hint on the HIV risk degree of each tweet in San Diego community.

4.4 Co-occurring HIV risk factors

Now that we understand that among HIV risk categories, the most pronounced are drug and homosexual relationship related terms, it is important to understand which of these categories have the highest likelihood of occurring together. In the following confusion matrix, we can get a clear understanding of co-occurrence characteristics of HIV risk terms.

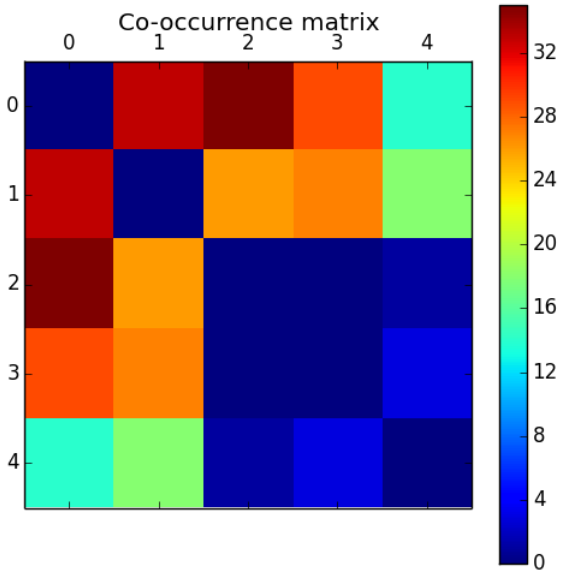

Figure 12: Degree distribution of HIV-risk tweets across users

The numbers in the rows and columns correspond to the following risk categories.

- 0 - Drugs
- 1 - Homosexual relationship terms
- 2 - Sexually Transmitted Infections
- 3 - Sex related terms
- 4 - Sex Venues in San Diego county

So having performed this extensive exploratory data analysis, now there is a concrete understanding of the interplay between various factors that make a tweet a HIV-risk tweet.

5. PREDICTIVE TASK CHOSEN

The goal of this project, as mentioned earlier, is to identify features that are most helpful in deciding whether a tweet show HIV risk or not, and finally come up with a good model for classifying an incoming tweet as a HIV-risk tweet. We compare two machine learning techniques for this.

1. Logistic Regression
2. Support Vector Machines

The tweets that show HIV risk fall under 5 different categories : Drug, Sex, Sex Venues, Homosexual behaviours, STI. The following 5 rules were used by the UCSD AVRC doctors to classify a tweet as HIV risk tweet or not.

1. Tweet which indicates the tweeter or someone around using drugs like meth, cocaine etc.
2. Tweet which indicates the tweeter or someone around being in a sex venue or gay bar known to be places of sex activities or drug exchanges.
3. Tweet which indicates the tweeter or someone around having a Sexually Transmitted Infection (STI).
4. Tweet which indicates the tweeter or someone around involved in a sexual activity.
5. Tweet which indicates the tweeter or someone around being MSM (Men who have Sex with Men).

Other two criteria which are complementary to the above 5 rules are that the tense of the tweet should be fairly present e.g. within past 24 hours. Secondly, the sentence should be non-news. This is how the labeled dataset was obtained. Once the model is built, it will be evaluated by testing it on a test set which is a subset of the labeled dataset.

Finally, the labeled dataset was split in the following way :
Training set size : 14000,
Validation set size : 6000,
Testing set size : 3205

5.1 Challenges

As mentioned in Ren et. al [1] , Twitter is notorious for being inert to traditional NLP techniques because of the following properties. Short text - The text is artificially limited to 140 characters. So tweeps (as they are called) use all sorts of emoticons and other hacky ways to convey more within the 140 character limit. Secondly, concept drift - Twitter is a social text stream. So the same content could mean different topics on different occasions.

6. RELATED WORKS

The work by Aramaki et. al. [2] tries to filter out negative influenza tweets from positive infuenza tweets by using a bag of words model and improving the performance of the classifier by focusing on words that are closer to the actual risk word (i.e. flu, influenza etc.). This paper uses a feature window size of 6 to the left and right of the risk word. But the authors did not talk about the exact mechanism used for extracting the features. They were able to get 89% correlation with their gold standard which are the IDSC reports.

Ren et. al [1] worked on classifying a short text document from a social text stream into multiple labels that can be organized in a hierarchy. They termed this classification model as Hierarchical Multi-label Classification (HMC). They expand the short document by linking to relevant Wikipedia articles and augmenting the text with titles from those Wikipedia articles. To handle the concept drift issue with social networks, they used a dynamic LDA model to infer global and local topics. Since our goal is to classify HIV risk tweets on a much shorter time frame we do not need to handle concept drift.

Malkani et. al. tried to evaluate the performance of SVM, Naive Bayes, Neural Network and Random Forests to classify tweets into attitudes and topics [3]. They were able to classify the sentiments and rationalize it based on the tweets that were generated during a Seahawks vs. Saints match. For achieving this, they discuss how they had engineered their features to include singletons, bigrams and then filtered the features based on frequency thresholds, mutual information and Chi-squared test.

Sun et. al [4] used graph based features like indegree, outdegree and a hybrid reputation parameter along with content based features to classify tweets as spam or not.

Sriram et. al [5] worked on classifying tweets into 5 different classes - news, events, opinions, deals and private messages. They showed that BOW-A (Bag of Words + Author) representation easily trumped over BOW (plain Bag Of Words) representation. Thus, this project also incorporates author information as features.

Banerjee et. al [6] showed that additional world knowledge helps in classifying short text like tweets. They were the first to show that extending the short document with sentences from Wikipedia helps to get better accuracy in text classification problems. They had evaluated their approach to a plain BOW representation.

Go et. al [7] from Stanford worked on extracting sentiments (positive or negative) from tweets. They used a concept called distant supervision where the dataset was considered "noisily" labeled by the emoticons present in the tweets. For instance, ":)" emoticon shows a positive sentiment while a ":(" emoticon present in the tweet shows a negative sentiment. They followed a similar approach as [3] in using unigrams, bigrams for engineering their features. They also used POS tags. Finally, they showed a comparative performance analysis on a bunch of text classification algorithms including SVM and Naive Bayes.

7. RESULTS

Feature Engineering was done as the first step. To begin with, a Bag-Of-Words (BOW) feature set was created. This feature set consisted of the 500 most frequently used words, 100 most frequently used hashtags and the 100 most frequently mentioned usernames. Then, all the stop words like "is", "they" etc. were removed so that the bag of words are more representative of the content of the tweets. The list of stop words from the Python library nltk was used for this purpose.

Then, it was time to leverage the domain expertise from the doctors at UCSD AVRC. Based on their interactions with the patients from the San Diego county, they helped create 5 lists of risk terms corresponding to each risk category mentioned earlier. The presence/absence of each of these risk terms will correspond to a 1 or 0 in the feature set. Finally, the feature set was augmented with BOW features from the user profile of the tweeter. This included the 500 most frequently occurring words in the "description" field of each user.

The following table summarizes the results. Since both the SVM and logistic regression models caused overfitting, both the models were regularized. The percentage corresponds to the error rate when the classifier was used.

Feature Set	**SVM**	**Logistic Regression**
Bag of Words	15.73%	15.72%
Stop word removal	12.9%	12.98%
Domain specific terms	11.37%	7.42%
Tweeter information	17.12%	15.23%

I believed that adding the user profile features would increase the accuracy of the classifier, but it did worse. To understand the cause, when I manually went through the user profiles of few of the highly influential twitter users who post HIV-risk tweets. Surprisingly, 8 out of every 10 users had an empty user description. This could be the reason for the worse performance after adding user profile features.

8. REFERENCES

1. Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn van Dolen, and Maarten de Rijke. 2014. âĂIJHierarchical multi-label classification of social text streamsâĂİ. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval(SIGIR '14). ACM, New York, NY, USA, 213-222. DOI=10.1145/2600428.2609595 http://doi.acm.org/10.1145/2600428.2609595

2. Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics

using Twitter." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

3. Zahan Malkani, and Evelyn Gillie. âĂIJSupervised Multi-class Classification of TweetsâĂİ. Part of CS229 course in Stanford University.

4. Bing Sun, and Chao Li. âĂIJTweets Filter and Topic ClassificationâĂİ

5. Sriram, Bharath, et al. "Short text classification in twitter to improve information filtering." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010. Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta.

6. "Clustering short texts using wikipedia." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

7. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford (2009): 1-12.

8. Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Preventive Medicine, 63, 112âĂŞ5. doi:10.1016/j.ypmed.2014.01.024