UNIVERSITY of York

This is a repository copy of *Pedagogical demonstration of twitter data analysis: A case study of world AIDS day, 2014*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/157265/

Version: Published Version

Article:
Fung, I.C.-H., Yin, J., Pressley, K.D. et al. (6 more authors) (2019) Pedagogical demonstration of twitter data analysis: A case study of world AIDS day, 2014. Scientific data. 84. ISSN 2052-4463

https://doi.org/10.3390/data4020084

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

data

MDPI

Article

Pedagogical Demonstration of Twitter Data Analysis: A Case Study of World AIDS Day, 2014

Isaac Chun-Hai Fung [1,*], Jingjing Yin [1], Keisha D. Pressley [1], Carmen H. Duke [1], Chen Mo [1], Hai Liang [2], King-Wa Fu [3], Zion Tsz Ho Tse [4] and Su-I Hou [5]

1 Department of Biostatistics, Epidemiology and Environmental Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30460, USA; jyin@georgiasouthern.edu (J.Y.); kp03651@georgiasouthern.edu (K.D.P.); cy00164@georgiasouthern.edu (C.H.D.); cm06957@georgiasouthern.edu (C.M.)

2 School of Journalism and Communication, Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; hailiang@cuhk.edu.hk

3 Journalism and Media Studies Centre, The University of Hong Kong, Hong Kong, China; kwfu@hku.hk

4 School of Electrical and Computer Engineering, College of Engineering, The University of Georgia, Athens, GA 30602, USA; ziontse@uga.edu

5 College of Community Innovation and Education, The University of Central Florida, Orlando, FL 32816, USA; su-i.hou@ucf.edu

* Correspondence: cfung@georgiasouthern.edu; Tel.: +1-912-478-5079

Received: 7 May 2019; Accepted: 5 June 2019; Published: 10 June 2019

Abstract: As a pedagogical demonstration of Twitter data analysis, a case study of HIV/AIDS-related tweets around World AIDS Day, 2014, was presented. This study examined if Twitter users from countries with various income levels responded differently to World AIDS Day. The performance of support vector machine (SVM) models as classifiers of relevant tweets was evaluated. A manual coding of 1,826 randomly sampled HIV/AIDS-related original tweets from November 30 through December 2, 2014 was completed. Logistic regression was applied to analyze the association between the World Bank-designated income level of users' self-reported countries and Twitter contents. To identify the optimal SVM model, 1278 (70%) of the 1826 sampled tweets were randomly selected as the training set, and 548 (30%) served as the test set. Another 180 tweets were separately sampled and coded as the held-out dataset. Compared with tweets from low-income countries, tweets from the Organization for Economic Cooperation and Development countries had 60% lower odds to mention epidemiology (adjusted odds ratio, aOR = 0.404; 95% CI: 0.166, 0.981) and three times the odds to mention compassion/support (aOR = 3.080; 95% CI: 1.179, 8.047). Tweets from lower-middle-income countries had 79% lower odds than tweets from low-income countries to mention HIV-affected sub-populations (aOR = 0.213; 95% CI: 0.068, 0.664). The optimal SVM model was able to identify relevant tweets from the held-out dataset of 180 tweets with an accuracy (F1 score) of 0.72. This study demonstrated how students can be taught to analyze Twitter data using manual coding, regression models, and SVM models.

Keywords: global health; health promotion; HIV/AIDS; social media; supervised machine learning; Twitter

1. Introduction

Globally, 36.9 million people were living with human immunodeficiency virus (HIV) and 2 million people became newly infected with HIV in 2014 [1]. The annual World AIDS Day (WAD) promotes HIV awareness and advocates for HIV prevention, treatment, and community support for people living with HIV. According to the Centers for Disease Control and Prevention (CDC), more than 1.1 million

people in the United States were estimated to be living with HIV in 2015, of whom approximately 15% were unaware of their HIV status [2]. Prevention of HIV infection and associated illnesses and deaths is one of the goals of the Healthy People 2020 initiatives in the United States [3].

Social media, such as Twitter and Facebook, has become increasingly popular as a data source and a tool in public health for both epidemiologic surveillance and communication surveillance [4]. Many public health agencies use social media to promote healthy lifestyles and disease prevention. For example, the CDC has specific Twitter profiles dedicated to HIV/AIDS prevention and control (@CDC_HIVAIDS and @talkHIV). Hence, prior studies examined using social media to deliver HIV prevention information [5]. Social media analysis revealed users' reactions to specific health promotion events [6]. Users from different countries might react to the same disease differently [7,8]. Research comparing Twitter contents of five different languages pertinent to the MERS outbreak in South Korea in 2015 found that users from different Asian countries had different concerns about the outbreak [7]. Research comparing English tweets with Chinese Weibo posts pertinent to Ebola also identified content topics specific to Chinese internet users [8]. The language used in tweets might reveal users' demographics, including their socio-economic status [9].

The potential of using Twitter as a tool for health communication and public health surveillance has long been recognized by researchers [4,10–12]. For example, researchers have explored Twitter's potential role in the surveillance of behaviors associated with increased risk of HIV infection [13]. The application of computer-enabled methods, such as keyword and hashtag analysis as well as supervised and unsupervised machine learning methods, may help improve digital surveillance [4,11] by scaling up content analysis of tweets pertinent to health topics, such as Ebola [8], pneumonia [14], polio [15], and Zika [16]. However, Twitter data analysis training for Master of Public Health (MPH) students largely remains an unmet need [17].

In the past few years, our team at Georgia Southern University has met our students' educational needs through various student projects [7,14,15,18–20]. Some projects focused on specific health topics, such as sentiment, contents, and retweets of vaccine-related tweets [7]. Others focused on a specific Twitter profile, such as the CDC's Office of Advanced Molecular Detection (@CDC_AMD) [18]. In many cases, mixed methods were used. Students were trained to manually code Twitter content. They were also trained to perform statistical analysis, such as regression models. Wherever appropriate, machine learning methods were introduced to the students.

In this paper, through two related MPH student projects, the readers are provided with a pedagogical demonstration on statistical methods that can be used to analyze manually coded HIV-related Twitter data. The research presented here was conducted as part of the educational experience for three graduate students, KDP, CHD, and CM, who were mentored by two faculty members, ICHF and JY, at Georgia Southern University in consultation with the other co-authors of this paper.

1.1. Part I

In the first student project (MPH capstone project), our pedagogical objective was to train MPH students how to perform manual coding of tweets and subsequent statistical analyses of association between meta-data variables and content variables. It was hypothesized that Twitter users residing in countries of different income levels might express different concerns on Twitter regarding HIV/AIDS. It should be noted that a Twitter user's location is self-reported. Additionally, the country's income level as defined by the World Bank was chosen as the basis of categorization. This was chosen for our pedagogical purpose as it was relatively objective and without too much controversy. Through analyzing Twitter data around WAD 2014, this hypothesis was tested:

H1: The contents of HIV-related tweets created around WAD 2014 (each content category as an outcome variable) vary by the self-reported location by country income level as defined by the World Bank.

1.2. Part II

In the second student project (a Special Topics class project), our pedagogical objective was to train MPH students in the application of an established supervised machine learning method to categorize the contents of HIV-related tweets created around WAD 2014 into, relevant, and, irrelevant, tweets. More specifically, it was aimed to train students on how to obtain the optimal support vector machine (SVM) model with the maximum F1 score and evaluate its performance using a manually coded held-out dataset. The F1 score is a measure of accuracy and is the harmonic mean of sensitivity and positive predictive value [21]. The goal for the analysis was to obtain benchmark data for future studies using automated methods to categorize tweets carrying the words, HIV, or, AIDS. into tweets that were genuinely relevant to HIV and those that were not.

Our study serves as a small step towards a better understanding of how future public health practitioners can be trained to analyze Twitter data to address research questions pertinent to health communication.

2. Data Description

Our data consisted of HIV/AIDS-related Twitter messages (i.e., tweets) surrounding WAD 2014. Data was collected from publicly available, user-generated contents from Twitter Advanced Search [22]. The advanced search allowed the use of several filters. Therefore, it was chosen to use three filters—query filter, language filter, and time filter, which fell under the sections, "All of these words", "Written in" and "From this date" respectively. With these filters, tweets matching the queried words, "HIV OR AIDS" in English within a three-day time frame from November 30, 2014, through December 2, 2014 (UTC +00:00) were gathered. People and place filters were not used, so any public user worldwide who posted a tweet in English about HIV or AIDS within the three-day time frame was included in the original sample. The original search yielded 184,349 original tweets (i.e., all retweets were excluded).

As there is a limit to the number of tweets that can be retrieved using the Twitter search Application Programming Interface (API), a python script (Supplementary Materials Python script S1) was written to retrieve the search results in March 2015 through web crawling. The contents of the tweets and their unique identification numbers (i.e., tweet IDs) were retrieved, and Twitter API was used to obtain additional information according to the tweet IDs. Information retrieved from the data collection included user ID number, username, user's location (available only if the user permitted), retweet count, message tweeted, time created, mention ID/IDs, and mention username. As geo-location data was only available for a minority of tweets—around 15% of tweets according to Liang, Sheng & Fu [23]—the self-reported locations of the Twitter users were used as a proxy. Among the 184,349 tweets, 131,407 of them had self-disclosed something in the location field, of which 103,928 (56.4%) were identifiable (unpublished analysis).

Detailed descriptions of data retrieval are presented in Appendix A.

3. Methods

3.1. Manual Coding

From the 184,349 original tweets, a 1% random sample was extracted (n = 1826) for manual coding. After reading around 200 randomly selected tweets from our dataset by the senior author, the coding scheme was developed. For the content analysis portion of the study, each of the sampled tweets was numerically coded based on 12 questions: 1, 2a–c, 3, 4a–b and 5a–e (Tables 1 and 2):

1. Language, written in English or not;
2. Reported location by:

    (a) Country income level as described by the World Bank yearly revised gross national income (GNI) per capita classifications [24],

(b) United States or not, and
(c) By state or territory if in the United States;

3. Specific keyword mention of, World AIDS Day, or, Red ribbon;
4. Mentions of:

(a) HIV/AIDS epidemiological content (e.g., incidence, prevalence, etc.),
(b) Sub-populations (e.g., age, race/ethnicity, gender, sexual orientation, drug use, other subgroup, etc.);

5. Mentions of:

(a) HIV/AIDS prevention and behavior content (e.g., abstinence, faithfulness to partner, condom use),
(b) HIV/AIDS testing,
(c) HIV/AIDS disclosure,
(d) Stigma/discrimination awareness, and
(e) HIV/AIDS compassion and support.

Table 1. Frequency table of 1824 manually coded tweets.

Question & Response	Frequency (%) [a]
1. Language: Was the tweet written in English?	
Yes	1769 (97.0)
2a. Self-Reported Location: Country income level as defined by the World Bank?	
Low-income countries (reference category)	42 (2.3)
Lower-middle-income countries	150 (8.2)
Upper-middle-income countries	98 (5.4)
High-income countries—non-OECD	15 (0.8)
OECD countries	677 (37.1)
Others	226 (12.4)
No information reported	616 (33.7)
3. Did the tweet mention WAD (or red ribbon)?	
Yes	795 (43.6)
4. HIV/AIDS Epidemiological Content	
4a HIV Epidemiology information mentioned?	
Yes (e.g., statistics provided, mentioned incidence/prevalence/epidemics, etc.)	142 (7.8)
4b HIV sub-populations mentioned?	
Yes	107 (5.9)
5. Content—HIV behaviors & prevention	
5a HIV/AIDS prevention information?	
Yes	30 (1.6)
5b HIV test(ing) mentioned?	
Yes	106 (5.8)
5c HIV disclosure mentioned?	
Yes	47 (2.6)
5d Stigma/discrimination awareness mentioned?	
Yes	103 (5.6)
5e HIV compassion and support mentioned?	
Yes (e.g., remembering the positives, support family, etc.)	542 (29.7)

[a] All the percentages included in this Table use 1824 as the denominator. IDU: injecting drug users; OECD: Organization for Economic Co-operation and Development

Table 2. The self-reported locations by state or territory of users from the United States (n = 453 tweets).

State/Territory	n (%)	State/Territory	n (%)
1. Alabama	5 (1.1)	31. New Mexico	2 (0.4)
2. Alaska	1 (0.2)	32. New York	76 (16.8)
3. Arizona	4 (0.9)	33. North Carolina	10 (2.2)
4. Arkansas	1 (0.2)	34. North Dakota	0 (0)
5. California	55 (12.1)	35. Ohio	13 (2.9)
6. Colorado	1 (0.2)	36. Oklahoma	1 (0.2)
7. Connecticut	1 (0.2)	37. Oregon	3 (0.7)
8. Delaware	0 (0)	38. Pennsylvania	13 (2.9)
9. Florida	27 (6.0)	39. Rhode Island	2 (0.4)
10. Georgia	13 (2.9)	40. South Carolina	10 (2.2)
11. Hawaii	7 (1.5)	41. South Dakota	0 (0)
12. Idaho	2 (0.4)	42. Tennessee	3 (0.7)
13. Illinois	19 (4.2)	43. Texas	26 (5.7)
14. Indiana	5 (1.1)	44. Utah	2 (0.4)
15. Iowa	4 (0.9)	45. Vermont	1 (0.2)
16. Kansas	1 (0.2)	46. Virginia	3 (0.7)
17. Kentucky	5 (1.1)	47. Washington	4 (0.9)
18. Louisiana	6 (1.3)	48. West Virginia	1 (0.2)
19. Maine	1 (0.2)	49. Wisconsin	3 (0.7)
20. Maryland	10 (2.2)	50. Wyoming	0 (0)
21.Massachusetts	11 (2.4)	51. Washington D.C.	16 (3.5)
22. Michigan	8 (1.8)	52. Puerto Rico	1 (0.2)
23. Minnesota	5 (1.1)	53. Guam	0 (0)
24. Mississippi	2 (0.4)	54. Other U.S. Territories	0 (0)
25. Missouri	2 (0.4)	55. U.S.A. Non-specific	46 (10.2)
26. Montana	1 (0.2)		
27. Nebraska	2 (0.4)		
28. Nevada	6 (1.3)		
29. New Hampshire	0 (0)		
30. New Jersey	12 (2.6)		

Results presented in Table 2 were based on coding to Questions 2b and 2c.

Using the coding scheme above, each tweet was grouped into respective categories. This study randomly selected 200 tweets (11%) to be coded by two coders independently. The inter-rater reliability was assessed by calculating the Cohen's kappa coefficient for the question variables (Appendix B. Table A1). Some of the values were low due to the extreme imbalance of the marginal totals. However, the observed proportion of agreement for each question was above 90%. Discrepancies were discussed and coding procedures were refined to address ambiguity in coding instructions and content meanings. The remaining tweets were divided and separately coded by the two coders.

Furthermore, a separate corpus of 180 tweets, randomly drawn from the original dataset, was coded by a third coder for the relevance of each tweet. This corpus was used as the held-out dataset for supervised machine learning.

3.2. Part I: Statistical Analysis

Statistical analysis was performed with R 2.15.0 to 3.2.1 [25]. Logistic regression was performed for response variables (i.e., content categories) that are binary. The primary predictor of interest in these regression models was the country's income level of self-reported locations, and all other related content categories were considered as confounders which needed to control for. In the regression analysis, Question 2b (United States versus non-United States) and Question 2c (if the United States, which state or territory) was not included due to their strong correlations with the country's income level. The latter also suffered from data sparsity (too few observations in one or more categories). This

study did not include Question 5a HIV prevention information from the regression models due to data sparsity. Regarding Question 4b, levels 1, 2, 3, 4, 5, and 6 were combined into one level for mention of HIV sub-populations (i.e., Yes, for any sub-population) so that there would be enough data in each category. In addition, due to strong multicollinearity between some of the content categories and self-reported locations, the stepwise model selection was used to determine the final regression models.

3.3. Part II: Support Vector Machine Model

SVM models are binary classifiers that can be trained, using a manually coded dataset, to separate tweets into two groups (relevant versus irrelevant). In Part II, all content categories were collapsed into one relevant category (i.e., all tweets coded with at least one content category) and one irrelevant category. The manually coded dataset of tweets in Part I (n = 1826) was used to create SVM models, with the aim of determining the optimal SVM model to test on a separate held-out dataset of 180 tweets. In this study, 1278 (70%) of the 1826 tweets in the manually coded sample were randomly selected as the training set for the SVM models, and the remaining 548 (30%) were used as the test set. The held-out dataset was separately randomly drawn from our data of 184,349 HIV/AIDS-related original tweets. The held-out data set was distinct from the sample of 1826 tweets in Part I, and the 180 tweets were manually coded as, relevant, or, irrelevant. The Twitter messages were preprocessed, with URL, digits, stop-words, and punctuation marks (except intra-word dashes) removed. Stemming was not performed to avoid the word, AIDS, being converted into, aid. Models were trained with variation in sparse term threshold within the document term matrices. Positive predictive value, sensitivity, specificity, and F1 scores were calculated for each training and test set. The optimal SVM model was identified based on the F1 score, which is the harmonic mean of sensitivity and positive predictive value, so a larger F1 score indicated the better prediction accuracy of the model. The trained optimal model was then used to predict whether the tweets in the held-out dataset of 180 tweets were relevant or not. SVM models were computed using R 3.2.1 to 3.2.3 [25].

3.4. Ethics Approval

The research presented herein was approved by the Georgia Southern University's Institutional Review Board, which determined it to be exempt from full review (H15083 and H15368).

4. Results

4.1. Part I: Statistical Analysis

Among our sample of 1826 tweets, two tweets were not properly coded by coder #2 for contents and were removed from subsequent analysis. Regarding the remaining 1824 tweets, self-reported locations of the Twitter users were not reported in 616 (33.4%) of the tweets.

From the 1208 tweets with self-reported locations of the Twitter users, the income levels of the (self-reported) countries of the Twitter users were as follows: Low-income countries (n = 42, 3.5%); lower-middle-income countries (n = 149, 12.3%); upper-middle-income countries (n = 98, 8.1%); high-income countries that are not members of the Organization for Economic Cooperation and Development (OECD) (n = 15, 1.2%); OECD member states (n = 683, 56.5%), and other locations (n = 221, 18.3%) (Table 1). The United States contributed to 453 (37.5%) of the 1,208 tweets with self-reported locations (Table 2), and the other 755 (62.5%) tweets were from non-United States locations.

Among the 1824 manually coded tweets, WAD was most commonly mentioned (n = 795, 43.2%), followed by support/compassion (n = 542, 29.4%), epidemiology (n = 142, 7.8%), any sub-population (n = 107, 5.9%), testing (n = 106, 5.8%), stigma/discrimination awareness (n = 103, 5.6%), disclosure (n = 47, 2.6%), and prevention information (n = 30, 1.6%) (Table 1).

Table 3 presents results from the logistic regression analysis. Compared with tweets from low income countries (reference category), tweets from OECD countries had 60% lower odds to mention HIV/AIDS epidemiology (adjusted odds ratio, AOR = 0.40; 95% confidence interval, 0.17, 0.98) after

controlling for mentions of sub-populations. Compared with tweets from low income countries, tweets from OECD countries had 3.08 times the odds of mentioning HIV/AIDS compassion and support (AOR = 3.08, 95% CI = 1.18, 8.05) after controlling for mentions of WAD, HIV/AIDS epidemiology, and HIV/AIDS testing. Tweets from lower middle income countries had 79% lower odds (AOR = 0.21; 95% CI, 0.07, 0.66) than low income countries to mention any of the sub-populations affected by HIV/AIDS after controlling for mentions of WAD and HIV/AIDS epidemiology. Our results supported our hypothesis that the contents of HIV-related tweets created around WAD 2014 vary by the self-reported location and by country income level as defined by the World Bank.

Table 3. Adjusted odds ratio of country income levels of self-reported locations and other variables as predictors of mentions of HIV/AIDS epidemiology information, mentions of sub-populations, and mentions of HIV/AIDS compassion and support in a step-wise multivariable regression analysis.

Question	Level of the Predictor Variable	Adjusted Odds Ratio	95% CI	*p*-Value
Outcome variable: Mentions of HIV/AIDS Epidemiology information				
Country income level of self-reported locations *	LI	reference	-	-
	LMI	0.804	0.300, 2.157	0.665
	UMI	0.548	0.183, 1.642	0.283
	HIC	0.974	0.166, 5.702	0.977
	OECD	0.404	0.166, 0.981	0.045
	Others	0.800	0.315, 2.031	0.639
Mentions of Sub-populations	Yes	7.226	4.408, 11.845	<0.001
Outcome variable: Mentions of Sub-populations				
Country income level of self-reported locations *	LI	reference	-	-
	LMI	0.213	0.068, 0.664	0.008
	UMI	0.523	0.175, 1.565	0.246
	HIC	0.298	0.030, 2.944	0.300
	OECD	0.424	0.176, 1.022	0.056
	Others	0.441	0.169, 1.148	0.093
Mentions of "World AIDS Day" or "Red Ribbon"	Yes	0.631	0.398, 0.998	0.049
Mentions of HIV/AIDS Epidemiology	Yes	6.856	4.168, 11.280	<0.001
Outcome variable: Mentions of HIV/AIDS compassion and support (Yes/No)				
Country income level of self-reported locations *	LI	reference	-	-
	LMI	2.126	0.765, 5.903	0.148
	UMI	1.782	0.611, 5.200	0.290
	HIC	3.885	0.895, 16.873	0.070
	OECD	3.080	1.179, 8.047	0.021
	Others	1.617	0.591, 4.426	0.349
Mentions of "World AIDS Day" or "Red Ribbon"	Yes	1.838	1.407, 2.401	<0.001
Mentions of HIV/AIDS Epidemiology	Yes	0.353	0.189, 0.658	0.001
Mentions of HIV/AIDS Testing	Yes	0.394	0.213, 0.731	0.003

CI, confidence interval. * Country income levels of self-reported locations: low-income countries as the reference category; LMI, lower-middle-income countries; UMI, upper-middle-income countries; HIC, high-income countries that are not members of the Organization for Economic Co-operation and Development; OECD, countries that are members of the Organization for Economic Co-operation and Development.

4.2. Part II: SVM Results

Five SVM models were created. The sparse term threshold varied from none to (n−10)/n, where n refers to the total number of terms (i.e., 4,963), giving 230 terms. Model variations, positive predictive values, and sensitivity are displayed in Table 4. The SVM model with a sparse term threshold of (n−5)/n gave the best F1 score (0.76) in the test set and was therefore used to test the manually coded held-out dataset of 180 tweets. The trained SVM model was able to predict the dataset with 77% sensitivity, a positive predictive score of 68%, and an F1 score of 0.72.

Table 4. Statistics of the performance of five support vector machine (SVM) models as described by sensitivity, positive predictive value and F1 score.

SVM Model	Sparse Term Threshold	Number of Terms	TP	TN	FP	FN	Specificity	Sensitivity	Positive Predictive Value	F1 Score
Training Set (n = 1278)										
A	0	4963	681	0	0	597	-	0.53	1	0.70
B	(n−3)/n	688	636	579	52	11	0.92	0.98	0.92	0.95
C	(n−5)/n	546	614	587	62	15	0.90	0.98	0.91	0.94
D	(n−7)/n	308	604	569	80	25	0.88	0.96	0.88	0.92
E	(n−10)/n	230	585	566	96	31	0.85	0.95	0.86	0.90
Test Set (n = 548)										
A	0	4963	Not applied to the test set because of poor performance in the training set							
B	(n−3)/n	688	216	176	63	93	0.74	0.70	0.77	0.73
C	(n−5)/n	546	235	168	56	89	0.75	0.73	0.81	0.76
D	(n−7)/n	308	222	174	61	91	0.74	0.71	0.78	0.74
E	(n−10)/n	230	211	190	75	72	0.72	0.75	0.74	0.74
Held-out Dataset (n = 180)										
C	(n−5)/n	546	68	60	32	20	0.65	0.77	0.68	0.72

FN: False Negative, FP: False Positive, TN, True Negative, TP: True Positive. The training set and test set were randomly selected from the corpus of 1826 tweets (70% and 30% respectively). F1 score is the harmonic mean of sensitivity and positive predictive value. Model C was selected because its F1 score on the test set was the highest among the five models. Model C was applied to the held-out dataset, which was a separate dataset of 180 manually coded tweets. Both the corpus of 1826 tweets and the corpus of 180 tweets were randomly selected from 184,349 HIV/AIDS-related original tweets from November 30 through December 2, 2014.

5. Discussion

This case study serves as a pedagogical demonstration of how public health graduate students could be taught statistical and text mining techniques for analyzing Twitter data. This study explored how contents of manually coded Twitter data pertinent to a specific health topic can be analyzed. In this case, our dataset contained tweets related to HIV/AIDS retrieved around WAD 2014. First, multivariable regression models were applied with the country income level of users as the main explanatory variable to explore differences in contents between users whose self-reported countries belong to different World Bank country income levels. Next, the SVM model was applied as a classifier for relevant and irrelevant tweets.

Our findings suggest a possibility of divergent interests between countries of different levels of economic development with regard to the types of HIV/AIDS-related information being shared on Twitter. These results echo observations that social media users reacted to infectious disease outbreaks differently between different communities or countries [7,8]. However, since location data is self-reported, and country-level income data does not apply to individuals (ecological fallacy), the authors caution against over-interpretation of the results of this exploratory study.

Our supervised machine learning analysis serves as a pilot for the use of SVM models to filter HIV-relevant tweets from irrelevant tweets in a corpus of tweets. The use of SVM models could reduce the amount of time needed to read through tweets and serve as the first sieve to pull relevant tweets for in-depth manual coding that will serve as the basis of qualitative analysis of the textual content of

Twitter data. The SVM method described here serves to facilitate, and not replace, qualitative studies that are tantamount to our understanding of social media contents.

5.1. Limitations

The dataset of this study is cross-sectional. Temporal variability of the data and causality between the predictive and outcome variables are beyond its scope. The three-day time frame was intentional to capture the responses to WAD. Hence, our results might not be applicable to other times of the year. Given the search terms, the sample of tweets was predominantly written in the English language, and thus our results cannot be generalized to users who wrote in other languages. Given the sample size for manual coding, the Twitter contents of users from individual countries could not be compared. Self-reported locations might not always accurately reflect the exact geographical location of the users. The geolocation data was not used because the proportion of tweets with geolocation data was low. The limitations of using the World Bank's categorization of country income levels of Twitter users' self-reported location as the explanatory variable are acknowledged. The diversity of cultures, healthcare systems, and prevalence of HIV/AIDS across countries (or territories) within the same country income level precluded the authors from drawing inferences between economic development and HIV-related concerns raised by Twitter users. Furthermore, it should be noted that a country's income level is not an indication of an individual's income level, and such an inference (which is known as ecological fallacy in epidemiology) should not be made. It is noted that in any country, Twitter users do not represent the general population. Finally, our study did not evaluate the quality (accuracy or reliability) of the information provided in the tweets. Only "mentions" were coded. Evaluation of the quality of HIV/AIDS-related information posted on Twitter is beyond the scope of this study.

5.2. Conclusions

To conclude, a pedagogical demonstration in public health Twitter data analysis using HIV-related tweets around WAD 2014 as a case study was presented. As social media data analysis becomes more mainstream in public health practice, training in big data analytical techniques will become more relevant to the education of our future public health practitioners.

Supplementary Materials: The following are available online at http://www.mdpi.com/2306-5729/4/2/84/s1, Python script S1.

Author Contributions: Conceptualization, I.C.-H.F., Z.T.H.T. and S.-I.H.; data curation, K.D.P. and C.H.D.; formal analysis, K.D.P., C.H.D. and C.M.; methodology, H.L.; project administration, I.C.-H.F.; resources, I.C.-H.F. and S.-I.H.; software, H.L.; supervision, I.C.-H.F. and J.Y.; writing—original draft, I.C.-H.F., J.Y., K.D.P., C.H.D. and H.L.; writing—review & editing, I.C.-H.F., J.Y., C.M., H.L., K.-W.F., Z.T.H.T. and S.-I.H.

Funding: This research received no external funding.

Acknowledgments: ICHF (15IPA1509134; 16IPA1609578) and ZTHT (16IPA1619505) received salary support from the Centers for Disease Control and Prevention (CDC). This paper is not related to their CDC-supported research. CDC has no role in the study design, data collection, data analysis, writing and submission of this paper. An early version of part I of this paper was submitted by KDP to ICHF as part of her Public Health Capstone Research Project (PUBH 7991), Spring 2015, titled "Examination of Self-reported Locations and Content of HIV/AIDS Twitter Data." An early version of part II of this paper was submitted by CHD to ICHF as part of her class project in Public Health Special Topics: Social Media and Health (PUBH 7090-A), Fall 2015. ICHF thanks Dr. Chung-Hong Chan (The University of Hong Kong and The University of Mannheim) who shared with him the text mining teaching materials for instructional use in his course. The authors thank Ogochukwu Nnaemeka Ezumba of The University of Georgia for being the second manual coder of tweets. ICHF, JY, KDP, CHD and CM serve as co-first authors.

Conflicts of Interest: The authors declare no conflict of interest. The CDC had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Technical Details of Twitter Data Retrieval

A.1. Original and Public Tweet Only

In Twitter Advanced Search, there is an option to include all retweets. This box was kept unchecked. Therefore, our data set included all the original tweets posted by any public users all around the world. According to Twitter, tweets posted by protected users are not searchable.

A.2. Filter Settings

- Language filter: English was selected.
- Query filter: Two keywords were used for our search: AIDS and HIV. These words were searched separately for technical considerations. Nevertheless, this was equivalent to search, AIDS OR HIV, at once.
- Time frame filter: This study included three days (2014-11-30, 2014-12-01, and 2014-12-02). The minimum unit is one day (not one hour). This study searched them day by day, and selected, from 2014-11-30 to 2014-12-01, in order to retrieve all tweets posted on 2014-11-30. The time zone used by Twitter search is UTC +00:00.

A.3. Searches

In summary, the authors searched 2 words (AIDS and HIV) in 3 days (2014-11-30, 2014-12-01, and 2014-12-02) separately. That means we searched 6 times (in parallel). The 6 URLs generated by the search interface are listed below:

- *AIDS, English, 2014-11-30:* https://twitter.com/search?f=realtime&q=%22AIDS%22%20lang%3Aen%20since%3A2014-11-30%20until%3A2014-12-01&src=typd
- *AIDS, English, 2014-12-01:* https://twitter.com/search?f=realtime&q=%22AIDS%22%20lang%3Aen%20since%3A2014-12-01%20until%3A2014-12-02&src=typd
- *AIDS, English, 2014-12-02:* https://twitter.com/search?f=realtime&q=%22AIDS%22%20lang%3Aen%20since%3A2014-12-02%20until%3A2014-12-03&src=typd
- *HIV, English, 2014-11-30:* https://twitter.com/search?f=realtime&q=%22HIV%22%20lang%3Aen%20since%3A2014-11-30%20until%3A2014-12-01&src=typd
- *HIV, English, 2014-12-01:* https://twitter.com/search?f=realtime&q=%22HIV%22%20lang%3Aen%20since%3A2014-12-01%20until%3A2014-12-02&src=typd
- *HIV, English, 2014-12-02:* https://twitter.com/search?f=realtime&q=%22HIV%22%20lang%3Aen%20since%3A2014-12-02%20until%3A2014-12-03&src=typd

Clicking these URLs will display the search results in the reverse chronical order (from the newest to the oldest). Normally, it displays 20 items (tweets) per page. When a user scrolls down, the page will continue loading 20 more items. Repeat scrolling will load all search results in a very long time.

A.4. Python Script.

A python script (Python script S1) was written to do this task automatically in March 2015. The python script is an automated browser, which imitates how people browse web pages. The script also recorded the unique tweet ID for each tweet. The authors did not scrape other information at this stage because it was usually unreliable. For example, the displayed time was in local form. Once the tweet IDs were obtained, other information was searched via the official API. Finally, 184,349 original tweets were obtained. For each tweet, the following fields were returned by the Twitter API:

- *user_id*: the unique id of the user who posted the tweet (please keep this confidential)
- *user_name*: the screen name of the user
- *user_location*: the self-reported location (this was used for geo-coding)

- *retweet_count*: the tweet has been retweeted how many times upon the data collection
- *txt*: the raw text of the tweet
- *created_at*: the time of posting
- *mentions_ids*: the unique ids of the mentioned users in txt, separated by comma
- *mentions_sn*: the screen names of the mentioned users in txt, separated by comma

Appendix B. Interrater Reliability

Table A1. Cohen's kappa for question 2a, 3, 4a-b and 5a-e. *.

Question & Response	Cohen's kappa
2a. Self-Reported Location: Country income level as defined by the World Bank?	0.65
3. Did the tweet mention WAD (or red ribbon)?	0.96
4. HIV/AIDS Epidemiological Content	
4a HIV Epidemiology information mentioned?	0.75
4b HIV sub-populations mentioned?	0.5
5. Content – HIV behaviors & prevention	
5a HIV/AIDS prevention information mentioned?	0.14
5b HIV test(ing) mentioned?	0.87
5c HIV disclosure mentioned?	0.55
5d Stigma/discrimination awareness mentioned?	0.67
5e HIV compassion and support mentioned?	0.37

* Cohen's kappa for each question was calculated using the original coding scheme. Categories for 4b were later collapsed during the analytics stage.

References

1. UNAIDS. Fact Sheet 2016. Available online: http://www.unaids.org/sites/default/files/media_asset/20150901_FactSheet_2015_en.pdf (accessed on 5 June 2019).
2. Centers for Disease Control and Prevention. HIV in the United States and Dependent Areas. Available online: http://www.cdc.gov/hiv/statistics/overview/ataglance.html (accessed on 5 June 2019).
3. Office of Disease Prevention and Health Promotion. Healthy People 2020—HIV. Available online: https://www.healthypeople.gov/2020/topics-objectives/topic/hiv/ (accessed on 5 June 2019).
4. Fung, I.C.-H.; Tse, Z.T.H.; Fu, K.W. The use of social media in public health surveillance. *West. Pac. Surveill. Response J.* **2015**, *6*, 3–6. [CrossRef] [PubMed]
5. Young, S.D.; Holloway, I.; Jaganath, D.; Rice, E.; Westmoreland, D.; Coates, T. Project HOPE: Online social network changes in an HIV prevention randomized controlled trial for African American and Latino men who have sex with men. *Am. J. Public Health* **2014**, *104*, 1707–1712. [CrossRef] [PubMed]
6. Fung, I.C.-H.; Cai, J.; Hao, Y.; Ying, Y.; Chan, B.S.B.; Tse, Z.T.H.; Fu, K.-W. Global Handwashing Day 2012: A qualitative content analysis of Chinese social media reaction to a health promotion event. *West. Pac. Surveill. Response J.* **2015**, *6*, 34–42. [CrossRef] [PubMed]
7. Blankenship, E.B.; Goff, M.E.; Yin, J.; Tse, Z.T.H.; Fu, K.W.; Liang, H.; Saroha, N.; Fung, I.C.-H. Sentiment, Contents, and Retweets: A Study of Two Vaccine-Related Twitter Datasets. *Perm. J.* **2018**, *22*, 17–138. [CrossRef] [PubMed]
8. Fung, I.C.-H.; Fu, K.W.; Chan, C.H.; Chan, B.S.; Cheung, C.N.; Abraham, T.; Tse, Z.T.H. Social Media's Initial Reaction to Information and Misinformation on Ebola, August 2014: Facts and Rumors. *Public Health Rep.* **2016**, *131*, 461–473. [CrossRef] [PubMed]
9. Preotiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; Aletras, N. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE* **2015**, *10*, e0138717. [CrossRef] [PubMed]
10. Sinnenberg, L.; Buttenheim, A.M.; Padrez, K.; Mancheno, C.; Ungar, L.; Merchant, R.M. Twitter as a Tool for Health Research: A Systematic Review. *Am. J. Public Health* **2017**, *107*, e1–e8. [CrossRef] [PubMed]
11. Jordan, S.E.; Hovet, S.E.; Fung, I.C.-H.; Liang, H.; Fu, K.-W.; Tse, Z.T.H. Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data* **2018**, *4*, 6. [CrossRef]

12. Tricco, A.C.; Zarin, W.; Lillie, E.; Jeblee, S.; Warren, R.; Khan, P.A.; Robson, R.; Pham, B.; Hirst, G.; Straus, S.E. Utility of social media and crowd-intelligence data for pharmacovigilance: A scoping review. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 38. [CrossRef] [PubMed]
13. Young, S.D.; Rivers, C.; Lewis, B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev. Med.* **2014**, *63*, 112–115. [CrossRef] [PubMed]
14. Adnan, M.M.; Yin, J.; Jackson, A.M.; Tse, Z.T.H.; Liang, H.; Fu, K.W.; Saroha, N.; Althouse, B.M.; Fung, I.C.-H. World Pneumonia Day 2011–2016: Twitter contents and retweets. *Int. Health* **2018**. [CrossRef] [PubMed]
15. Schaible, B.J.; Snook, K.R.; Yin, J.; Jackson, A.M.; Ahweyevu, J.O.; Chong, M.; Tse, Z.T.H.; Liang, H.; Fu, K.W.; Fung, I.C.-H. Twitter conversations and English news media reports on poliomyelitis in five different countries, January 2014 to April 2015. *Perm. J.* **2019**, *23*, 18–181.
16. Fu, K.W.; Liang, H.; Saroha, N.; Tse, Z.T.H.; Ip, P.; Fung, I.C.-H. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *Am. J. Infect. Control* **2016**, *44*, 1700–1702. [CrossRef] [PubMed]
17. Fung, I.C.-H.; Tse, Z.T.H.; Fu, K.W. Converting Big Data into public health. *Science* **2015**, *347*, 620. [CrossRef] [PubMed]
18. Fung, I.C.-H.; Jackson, A.M.; Mullican, L.A.; Blankenship, E.B.; Goff, M.E.; Guinn, A.J.; Saroha, N.; Tse, Z.T.H. Contents, Followers, and Retweets of the Centers for Disease Control and Prevention's Office of Advanced Molecular Detection (@CDC_AMD) Twitter Profile: Cross-Sectional Study. *JMIR Public Health Surveill.* **2018**, *4*, e33. [CrossRef]
19. Jackson, A.M.; Mullican, L.A.; Yin, J.; Tse, Z.T.H.; Liang, H.; Fu, K.W.; Ahweyevu, J.O.; Jenkins III, J.J.; Saroha, N.; Fung, I.C.-H. #CDCGrandRounds and #VitalSigns: A Twitter Analysis. *Ann. Glob. Health* **2018**, *84*, 710–716. [PubMed]
20. Fung, I.C.-H.; Jackson, A.M.; Ahweyevu, J.O.; Grizzle, J.H.; Yin, J.; Tse, Z.T.H.; Liang, H.; Sekandi, J.N.; Fu, K.W. #Globalhealth Twitter Conversations on #Malaria, #HIV, #TB, #NCDS, and #NTDS: A Cross-Sectional Analysis. *Ann. Glob. Health* **2017**, *83*, 682–690. [PubMed]
21. Sasaki, Y. The Truth of the F-Measure. 2007. Available online: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf (accessed on 5 June 2019).
22. Twitter Advanced Search. Available online: https://twitter.com/search-advanced (accessed on 5 June 2019).
23. Liang, H.; Shen, F.; Fu, K.-W. Privacy protection and self-disclosure across societies: A study of global Twitter users. *New Media Soc.* **2017**, *19*, 1476–1497. [CrossRef]
24. World Bank Country and Lending Groups. Available online: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed on 5 June 2019).
25. R: A Language and Environment for Statistical Computing. Available online: http://www.R-project.org (accessed on 5 June 2019).

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).