University of South Carolina
Scholar Commons

Faculty Publications | Information Science, School of

2021

Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning

Amir Karami
University of South Carolina, karami@mailbox.sc.edu

M. Lundy

F. Webb

H. R. Boyajieff

M. Zhu

See next page for additional authors

Follow this and additional works at: https://scholarcommons.sc.edu/libsci_facpub

 Part of the Library and Information Science Commons

Publication Info

Electronics, Volume 10, Issue 15, 2021.
© 2021 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.

This Article is brought to you by the Information Science, School of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.

Author(s)
Amir Karami, M. Lundy, F. Webb, H. R. Boyajieff, M. Zhu, and D. Lee

This article is available at Scholar Commons: https://scholarcommons.sc.edu/libsci_facpub/393

electronics

MDPI

Article

Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning

Amir Karami [1,*], Morgan Lundy [2], Frank Webb [3], Hannah R. Boyajieff [4], Michael Zhu [5] and Dorathea Lee [6]

[1] School of Information Science, University of South Carolina, Columbia, SC 29208, USA
[2] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA; melundy2@illinois.edu
[3] Department of Computational Social Science, George Mason University, Fairfax, VA 22030, USA; fwebb2@masonlive.gmu.edu
[4] Darla Moore School of Business, University of South Carolina, Columbia, SC 29208, USA; boyajieh@email.sc.edu
[5] Department of Psychology, University of South Carolina, Columbia, SC 29208, USA; mz3@email.sc.edu
[6] Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC 29208, USA; dorathea@email.sc.edu
* Correspondence: karami@mailbox.sc.edu

Abstract: Privacy needs and stigma pose significant barriers to lesbian, gay, bisexual, and transgender (LGBT) people sharing information related to their identities in traditional settings and research methods such as surveys and interviews. Fortunately, social media facilitates people's belonging to and exchanging information within online LGBT communities. Compared to heterosexual respondents, LGBT users are also more likely to have accounts on social media websites and access social media daily. However, the current relevant LGBT studies on social media are not efficient or assume that any accounts that utilize LGBT-related words in their profile belong to individuals who identify as LGBT. Our human coding of over 16,000 accounts instead proposes the following three categories of LGBT Twitter users: individual, sexual worker/porn, and organization. This research develops a machine learning classifier based on the profile and bio features of these Twitter accounts. To have an efficient and effective process, we use a feature selection method to reduce the number of features and improve the classifier's performance. Our approach achieves a promising result with around 88% accuracy. We also develop statistical analyses to compare the three categories based on the average weight of top features.

Keywords: LGBT; social media; machine learning; Twitter

check for updates

Citation: Karami, A.; Lundy, M.; Webb, F.; Boyajieff, H.R.; Zhu, M.; Lee, D. Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning. *Electronics* **2021**, *10*, 1822. https://doi.org/10.3390/electronics10151822

Academic Editors: Amir H. Gandomi, Fang Chen and Laith Abualigah

Received: 20 June 2021
Accepted: 27 July 2021
Published: 29 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There are more than 11 million lesbian, gay, bisexual, and transgender (LGBT) adults in the United States (U.S.) [1]. LGBT populations face social stigma and additional discrimination-imposed challenges such as a higher rates of HIV and depression than their heterosexual and cisgender peers [2]. To address LGBT issues and provide better service to this community, the first step is identifying those issues. However, traditional surveys and other research approaches such as focus groups are expensive and time-consuming, address limited issues, and obtain small-scale data.

Social media has become a mainstream channel of communication and has grown in popularity. Social media facilitates people's belonging to and exchanging information within LGBT communities by allowing users to transcend geographic barriers in online spaces with the limited risk of being "outed" [3]. Compared to heterosexual respondents, LGBT users are more likely to have accounts on social media websites, access social media daily, and make frequent use of the internet [4].

According to a survey, 80% of LGBT Americans use social networking websites, and about four in ten LGBT adults have revealed their sexual orientation or gender identity on

social networking sites [5]. These statistics show that identifying and exploring millions of LGBT users and their discussions on social media could lead to revolutionary new ways of collecting and analyzing data about LGBT people, impacting a range of disciplines, including health, informatics, sociology, and psychology.

Privacy and stigma pose significant barriers to LGBT people sharing information related to these identities [6]. Therefore, social media data may provide unique perspectives on LGBT issues that are not shared in these other settings. Among different social media platforms, Twitter offers Application Programming Interfaces (API) to collect large-scale datasets [7,8]. Due to its APIs and millions of users, many studies have used Twitter data to examine phenomena of interests in different applications such as health [9–14], politics [15–17], social issues such as sexual harassment [18], and disaster analysis [19–21].

Valuable social media research has been implemented about the LGBT community on social media, such as defining online identity [22,23], exploring the societal perception about the LGBT community [24], investigating transgender adolescents' uses of social media for social support [25], analyzing how LGBT parents navigate their online environments for advocacy, privacy, and disclosure [26], comparing characteristics of research participants recruited via in-person intercept interviews in LGBT social venues and targeted social media ads [27], and identifying health issues of LGBT users on social media [28–30]. Studies have also been conducted exploring the visibility and participation of LGBT users online [31], the correlation between psychological wellbeing and social media dependency [32,33], LGBT hospital care evaluation based on social media posts [34], gender transition sentiment (mental health) [35], the health and social needs of transgender people [36], sexual health promotion [37–44], and the intervention and recruitment of research participants [45–54].

The above studies have utilized three approaches to identify LGBT users on social media. The first one places calls for LGBT participants, such as [45]. The second approach manually finds profiles related to self-identified individuals, such as [30]. The last approach is to identify profiles containing LGBT-related words, such as [35]. These approaches have limitations. The first and second approaches are time-consuming and labor-intensive. The limitation of the third approach is that users who utilize LGBT-related words in their profile are not necessarily LGBT individual users and can belong to other types of users, such as organizations. Similar to our research, some studies have developed classifiers to identify the gender or age of users in text data [55] and in social media such as Twitter [56,57], Sina Weibo [58], Facebook [59], and Netlog [60]. These studies have utilized different features such as the number of pronouns in social media profile and bio to develop binary classifiers, but there is no research on identifying the category of LGBT users.

In sum, social media has provided a great opportunity for both LGBT users to overcome the relevant privacy and stigma issues and researchers to study LGBT population. However, there is a need to develop efficient and effective automated methods to categorize LGBT users using machine learning. To addresses the limitations, first, this research develops a robust codebook to characterize users in community-informed ways beyond just searching by keywords, to provide the most accurate data to train a machine learning model. Then, this study offers a classifier to automatically categorize LGBT users to facilitate future relevant studies. This paper has multiple contributions and implications, as follows:

- This paper offers a codebook to manually categorize LGBT users.
- The prediction approach is an important step toward categorizing LGBT users by developing a machine learning classifier.
- Methodologically, our approach can be reused in predicting not only LGBT users but also other minorities.
- While this research uses Twitter data, the proposed approach and features can be adopted for other possible social media platforms.
- The approach of this paper can be used to identify and filter out adult content.

- This research can be used by researchers to understand social media activities and concerns (e.g., health issues) of LGBT individuals.
- This study can also be utilized by researchers to explore the social media strategies of LGBT organizations and identify best practices to promote social good for the LGBT population.

2. Materials and Methods

The methodology of this paper has five components, including data acquisition, data annotation, classification, evaluation, and statistical analysis (Figure 1).

Figure 1. Research Methodology Components.

2.1. Data Acquisition

Twitter data were chosen for this project due to the hesitations many LGBT community members have about reporting their identities in official studies or in medical settings. Choosing Twitter data also allows for a broader reach within the community than is possible in a survey approach. Survey or focus-group based research on queer issues may also be heavily siloed, while Twitter data offer a broader view available at scale.

Twitter, a massively popular American microblogging and social networking platform launched in 2006, allows users to post short messages or "tweets" and interact with other users' tweets by liking or retweeting. Users choose to "follow" other users whose content they wish to view and can choose to only allow certain other users to follow their account. Twitter is a social media platform that provides us with a large-scale dataset to classify LGBT users.

This paper categorizes Twitter users utilizing LGBT-related words in their profiles. Profiles were identified using the followerwonk platform (https://followerwonk.com/bio (accessed on 15 June 2019)) to obtain Twitter profiles containing "lesbian", "gay", "bisexual", "bi", "transgender", "trans man", and "trans woman" users in the U.S. and in each state, and only profiles that had at least 50 followers and 50 tweets to focus on active users. This process offered 42,644 profiles. After removing duplicate profiles, we found 38,978 unique profiles.

We recognize that the topic of this paper is a sensitive area presenting ethical challenges. To address these challenges, we include a self-reflexivity statement. First, we use publicly accessible Twitter data without any interaction with the users, our work is exempt from the institutional review board (IRB) review. However, we took great care in data collection and analysis and presenting results by not disclosing personally identifiable information. Second, to incorporate sensitivity in this paper, some of the coauthors belong to the LGBT community.

2.2. Data Annotation

In order to accurately categorize LGBT users, high quality data from users who self-identify as LGBT in the United States are needed. Where previous work in the field has taken more simplistic approaches to gather profiles belonging to the community by simply including all profiles with mentions of LGBT terms, this results in low quality data due to the inclusion of accounts professing support as allies and automated accounts that post primarily pornographic material. This research could be used to automate the process of future classification and could serve as a repository for a number of future academic studies into many other aspects of LGBT social media activities.

The annotation approach and codebook were developed iteratively, and responsively to both community rhetoric and the intricacies, twists, and unexpected challenges of mining social media data. Using a human-centered approach, a codebook was developed to reflect the most complexity possible when labeling the accounts of users, while still creating

disjoint sets. The final codebook was then applied to all collected user accounts by two coders independently for intercoder reliability.

The two authors independently coded and discussed 500 randomly selected profiles from the 38,978 unique profiles. Due to the nature of the internet and social media at large, searching for profiles with LGBT-related words in Twitter bios returns a fairly high percentage of results with primarily pornographic material, which may or may not be posted by "bots". Organizations were also classified separately, as they do not reflect individual experiences. Discrepancies were addressed by a third coder. The initial coding process offered three categories, including individual, porn/sex worker, and organization accounts. Coders needed to answer the following two questions for each account:

Q1: Is the account useable for this research? This yes/no question excluded the following accounts:

- Non-U.S. accounts where their bio information does not show a location in the U.S.;
- Non-English accounts that posted mostly non-English tweets;
- Inactive accounts that have not been active since 2017;
- Private and suspended accounts;
- Automated accounts that posted an unusual number of tweets, retweets, and likes, had a very low rate of followers to followings, and did not have an image. We also used Botometer (https://botometer.osome.iu.edu/ (accessed on 15 June 2019)) to identify automated accounts [61].

Q2: What is the category of the account? To address this question, coders used the following definition to assign one of the categories:

- Individual accounts are controlled by a single person.
- Sex Worker/Porn accounts are involved in the production of professional pornography both on and off screen, those engaged in prostitution and escort services, erotic dancers, fetish models, and amateur individuals using webcam sites, amateur porn sites, or pay-gated platforms to profit off of self-made content, and accounts that retweet primarily pornographic material and/or post their own nude photographs or moving images.
- Organization accounts are managed by a group or an organization representing more than one person.

After completing the coding, we applied Cohen's κ to determine the agreement between the two coders. There were substantial agreements for Q1 ($\kappa = 0.7862$) and Q2 ($\kappa = 0.7544$).

2.3. Classification

Our next goal centers around inferring the category of the collected Twitter users automatically. We draw on Twitter account information to build a machine learning classifier. This paper follows the automated framework in Figure 2 to categorize LGBT users on Twitter.

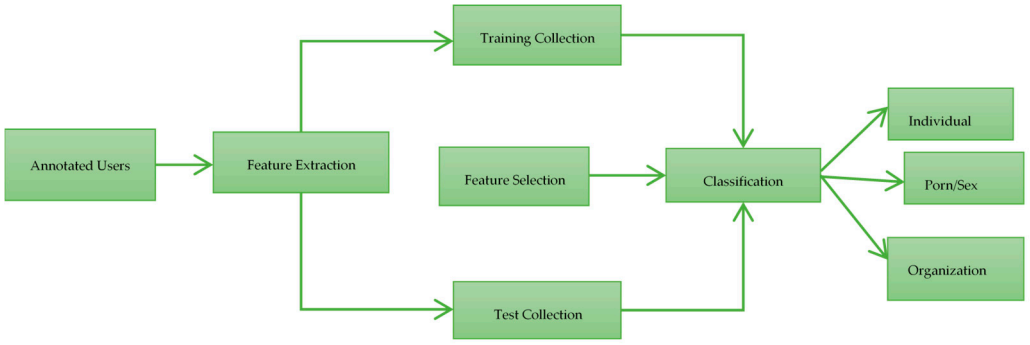

Figure 2. Classification Framework.

This step includes developing algorithms to assign a set of users $U = \{u_1, u_2, \ldots, u_k\}$ to known classes. The classification can be described as the prediction of the category of each user (u_i). The following classifier algorithm (a) assigns a class (c) to each user in Equation (1):

$$a : u_i \rightarrow \{c_1, c_2, \ldots, c_m\} \quad (1)$$

In this research, there are three classes (m = 3), including individual, sex worker/porn, and organization. To classify each user, the input of each classifier is a set of n features, $F = \{f_1, f_2, \ldots, f_n\}$. This research examines the following two types of features: bio and profile features. To predict the category of each user, we use the following two main approaches [62]: traditional methods including NaiveBayes, BayesNet, Random Forest, J48, and Support Vector Machines (SVM) and deep learning using Convolutional Neural Network (CNN). These methods are among high-performance classification algorithms [63–68]. CNN is of the popular deep learning methods and has been used for different classification tasks [62,69–71]. The rest of the classifiers are traditional machine learning methods using for a wide range of applications such as spam detection [72,73] and document classification [74]. We transform the information of Twitter accounts into a set of features. The focus of this study is on the features displayed on Twitter accounts. These features illustrate information about users and their activities. Table 1 shows the definition of Twitter terms.

Table 1. Twitter Terminology.

Term	Definition
Account's Age	The length of time that a Twitter account has been created.
Bio	A short summary (up to 160 characters) about a user in their profile.
Like (Favorite)	Showing appreciation of a tweet by clicking on the like tab.
Followers	Twitter accounts that follow updates of a Twitter account.
Followings	Twitter accounts that are followed by a Twitter account.
Screen Name	The name displayed in the profile to show a personal identifier.
Tweet	A status update of a user containing up to 280 characters.
Username	The name to help identify a user using @, such as @TheEllenShow.

This paper uses features in the LGBT Twitter accounts and builds a feature vector for each account, which are briefly described below.

- 1288 bio features
  - Frequency of each word in bio
  - The number of words in bio
- 81 profile features
  - The age of each Twitter account (account's age)
  - Total number of tweets (#tweets)
  - The number of tweets per year (#tweets/year)
  - The number of likes per year (#likes/year)
  - The number of followers (#followers)
  - The number of followings (#followings)
  - The rate of followers to following (#followers/#followers)
  - Frequency of letters (A–Z) and numbers (0–9) in the username
  - Frequency of letters (A–Z) and numbers (0–9) in the screen name
  - The username's length
  - The screen name's length

This study uses the χ^2 value, which is one of the effective feature selection methods [75], to measure the discriminative power of features for ranking the impact of the

different number of features on the performance of classification methods. The χ^2 value assists in identifying the best number of features showing the best classification performance.

2.4. Evaluation

We examine the performance of the six algorithms to find which classifier performs better with the bio and profile features. To evaluate the performance of classifiers, we use some measures based on the confusion matrix. The following confusion matrix represents a binary classification example that can be extended to more than two categories:

		Predicted	
		Category 1	Category 2
Actual	Category 1	True Positive (TP)	False Positive (FP)
	Category 2	False Negative (FN)	True Negative (TN)

While TP and TN are correctly identified and misidentified reports, respectively, FP and FN are incorrectly identified and misidentified reports, respectively. We utilized precision (P), recall (R), the area under the ROC curve (AUC), and accuracy (ACC) based on the following definitions:

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FP} + \text{TN}}$$

ROC finds the tradeoff FP and TP by plotting FP on the X-axis and TP on the Y-axis; the closer to the upper left indicates better performance. Then, we computed the chi-square to rank and find the top features. In order to determine the category of each user, we adopted the six classification algorithms using 5-fold cross-validation, in which the data are broken into five subsets, and the holdout method is repeated five times. Each time, one of the three subsets is used as the test set, and the other four subsets are used as the training set.

2.5. Statistical Analysis

To compare individual, porn/sexual worker, and organization accounts based on the mean value of the top features identified in the previous step, we utilized an analysis of variance (ANOVA), which tests whether the weight of features is different for the three account's types. We used the value of the top features as the dependent variable. After we found a significant difference (p-value ≤ 0.05), we used Tukey's multiple comparison test [76] to find which of the means differ significantly from others. To control familywise errors, we used the false discovery rate (FDR) method [77] that reduces not only false positives but also false negatives [78]. We also utilized the absolute effect size using Cohen's d to identify the magnitude of the differences. We used the following classification index to interpret effect sizes: very small (d = 0.01), small (d = 0.2), medium (d = 0.5), large (d = 0.8), very large (d = 1.2), and huge (d = 2.0) [79].

3. Results

The manual coding process offered 16,241 users, including 12,488 (76.89%) individual, 2282 (14.05%) porn/sexual work, and 1471 (9.06%) organization accounts. In total, we obtained 1369 features. We tested the performance of the six classifiers developed in Weka (https://www.cs.waikato.ac.nz/ml/weka/ (accessed on 15 April 2021) with the five-cross validation methods. To ensure the comparability between the classifiers, we used the standard parameters. Out of the six classifiers, we found BayesNet produced higher accuracy and AUC than the rest of the algorithms (Figure 3). The BayesNet algorithm

performed significantly better than the baseline accuracy of 0.7689, which was based on using the algorithm ZeroR relying on the target and ignores all predictors.

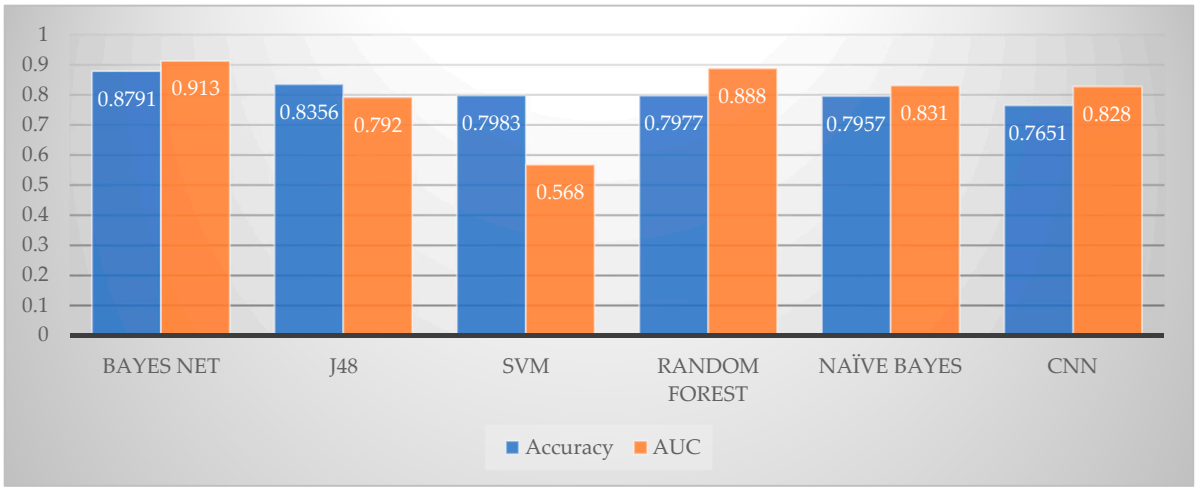

Figure 3. Classification performance of six algorithms using 1369 features.

We found that finding the optimum number of features can improve the classification performance, which offers a time-saving and cost-efficient system. Therefore, we have examined a different number of features. The optimum number of features was 399 (Figure 4).

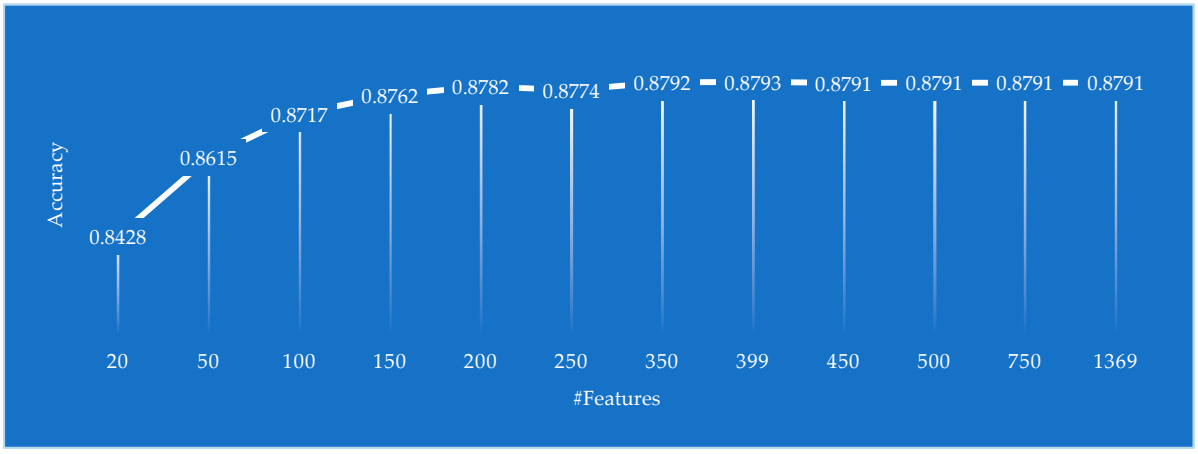

Figure 4. Accuracy of different number of features.

Table 2 shows the accuracy performance of BayesNet algorithms using the profile, bio, profile and bio, and top 399 profile and bio features. This table had three outcomes. First, the profile or bio features could identify the three classes with more than 80% accuracy. Second, the combination of profile and bio improved the performance of the classifier. Third, reducing the number of features enhanced the accuracy of BayesNet.

Table 2. Classification Performance of BayesNet.

Model	#Features	Acc	AUC	Precision	Recall
Profile	81	0.8089	0.811	0.790	0.809
Bio	1288	0.8596	0.874	0.851	0.860
Profile and Bio	1369	0.8791	0.913	0.873	0.879
Profile and Bio	399	0.8793	0.913	0.873	0.879

Table 3 summarizes the performance metrics of NaiveBayes with 399 features, where we found that the classifier was reasonably stable (SD $\leq$ 0.006 and CV $\leq$ 0.01). CV represents the coefficient of variation measured using $\frac{\text{Standard Deviation}}{\text{Mean}}$.

Table 3. Detailed Classification Performance using NaiveBayes and top 399 Features.

Metric	Min	Max	Mean	SD	CV
Accuracy	0.8750	0.8824	0.8793	0.003	0.004
AUC	0.905	0.922	0.913	0.006	0.007
Precision	0.869	0.876	0.873	0.003	0.004
Recall	0.875	0.882	0.879	0.003	0.004

Among the top 399 features, the number of bio features is more than the number of profile features (Figure 5). We found that 321 (24.9%) out of the total 1288 bio features and 78 (96.3%) out of the total 81 profile features are among the top 399 features. This means that while the profile features represent 20% of the top features, most profile features are among the top features. Among the top 50 features, the number of profile features is more than the number of bio features.

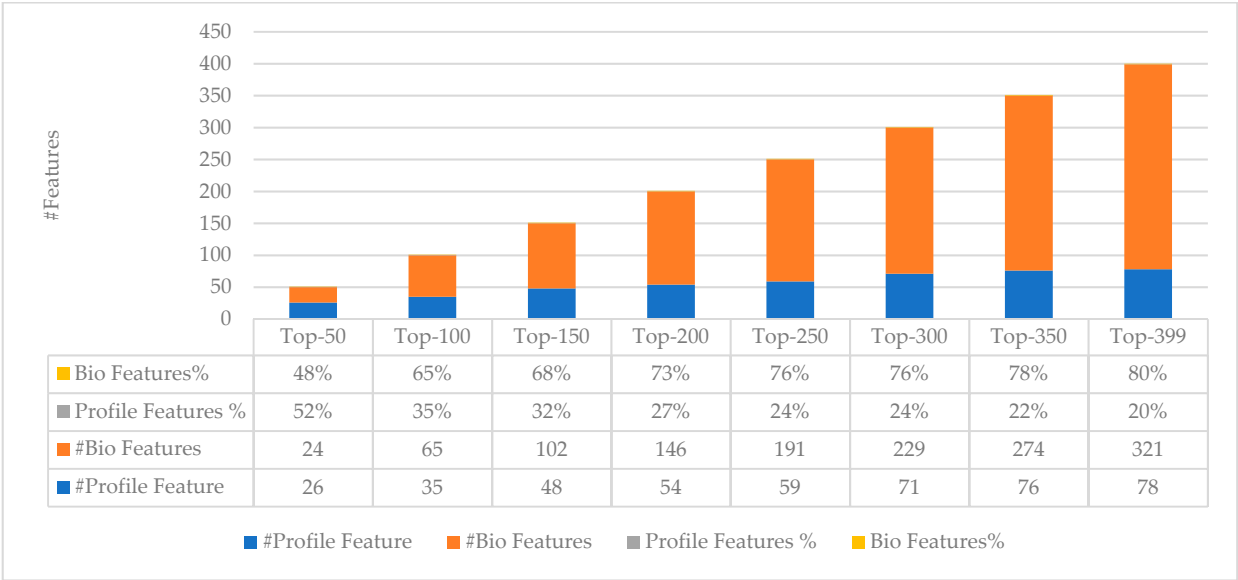

Figure 5. The number and the rate of bio and profile features among the top 399 features.

Table 4 shows the top 20 features that assisted in classifying users. Out of the top 20 features, 9 (45%) and 11 (55%) features were related to profile and bio categories, respectively. The bio features include the #words in the bio and the frequency of words in the bio of Twitter accounts, including bisexual, transgender, community, nsfw, porn, organization, LGBT, allies, event, and men. Among these words, nsfw and porn are used more by SWP accounts, and the rest of words are utilized more by organization accounts. The rest of the top 20 features are related to the profile category, including the #likes/year, the #followers/#followings, the account's age, the #tweets/year, the #tweets, the letter g in the screen name, the #followers, the username's length, and the #followings.

Table 4. Statistical Analysis of Top 20 Features of Classifiers (Ind: Individual; Org: Organization; SWP: Sexual Worker/Porn; *: p-value ≤ 0.05; NS: Not Significant).

Feature	Category	ANOVA + Tukey Multiple Comparison		
		SWP vs. Org	**Ind vs. Org**	**Ind vs. SWP**
#Likes/Year	Profile	* SWP > Org	* Ind > Org	NS
Bisexual	Bio	* SWP < Org	* Ind < Org	NS
Transgender	Bio	* SWP < Org	* Ind < Org	NS
Community	Bio	* SWP < Org	* Ind < Org	NS
#Followers/#Followings	Profile	NS	* Ind < Org	* Ind < SWP
Nsfw	Bio	* SWP > Org	NS	* Ind < SWP
Account's Age	Profile	* SWP < Org	* Ind < Org	* Ind > SWP
#Tweets/Year	Profile	* SWP > Org	* Ind > Org	NS
#Tweets	Profile	* SWP > Org	* Ind > Org	* Ind > SWP
Porn	Bio	* SWP > Org	* Ind < Org	* Ind < SWP
Letter G in Screen Name	Profile	* SWP < Org	* Ind < Org	* Ind < SWP
Organization	Bio	* SWP < Org	* Ind < Org	NS
#Followers	Profile	NS	* Ind < Org	* Ind < SWP
LGBT	Bio	* SWP < Org	* Ind < Org	* Ind > SWP
Username's Length	Profile	* SWP < Org	* Ind < Org	NS
Allies	Bio	* SWP < Org	* Ind < Org	NS
#Words in Bio	Bio	* SWP < Org	* Ind < Org	* Ind < SWP
Events	Bio	* SWP < Org	* Ind < Org	NS
#Followings	Profile	NS	NS	* Ind < SWP
Men	Bio	* SWP < Org	* Ind < Org	* Ind < SWP

Our statistical analysis shows that there were 46 (out of 60) significant differences. For example, the number of tweets is higher for individual (Ind) accounts than sexual worker/porn (SWP) and organization (Org) accounts (Table 4). We found the following findings:

- There was no significant difference between SWP and Org accounts across three features, including #followers/followings, #followers, and #followings. Compared to Org accounts, SWP accounts had a higher #likes/year, #tweets/year, and #tweets and used nsfw (not safe for work) and pornographic words in their bio more. The value of the rest of the features was higher for Org accounts than SWP ones. In sum, we found three NS, five SWP > Org, and twelve SWP < Org comparisons.
- There was no significant difference between Ind and Org accounts across two features, including #followings and nsfw. Compared to Org accounts, Ind accounts had a higher #likes/year, #tweets/year, and #tweets. The value of the rest of the features was higher for Org accounts than Ind ones. In total, this research identified two NS, three Ind > Org, and fifteen Ind < Org comparisons.
- There was no significant difference between Ind and SPW accounts across nine features, including #likes/year, #tweets/year, and the length of the username, and using the words bisexual, transgender, community, organization, and allies in their bio. Compared to SPW accounts, Ind accounts had a higher account age, #tweets, and used the acronym LGBT more. The value of the rest of features was higher for SPW accounts than Ind ones. In total, this research identified nine NS, three Ind > SWP, and eight Ind < SWP comparisons.

- There is a significant difference between the three categories based on the following features: the account's age; the number of tweets; using porn, LGBT, and men words in the bio; using G in screen name; and the number of words in the bio.
- The effect size analysis illustrates that the 46 significant differences were not trivial, including 6 very small, 18 small, 14 medium, 6 large, and 2 very large effect sizes (Table 5). The maximum difference was between individual and organization accounts with 18 (90%) significant differences, and the minimum difference was between the individual and sexual worker/porn accounts with 11 (55%) significant differences out of 20 comparisons. The effect size analysis also confirmed that the magnitude of significant differences is considerable.

Table 5. Effect Size Analysis of Top 20 Features (Ind: Individual; Org: Organization; SWP: Sexual Worker/Porn).

Feature	Cohen's d			Effect Size		
	SWP vs. Org	Ind vs. Org	Ind vs. SWP	SWP vs. Org	Ind vs. Org	Ind vs. SWP
#Likes/Year	0.58	0.43	NS	Medium	Small	NS
Bisexual	0.89	1.66	NS	Large	Very Large	NS
Transgender	0.86	1.58	NS	Large	Very Large	NS
Community	0.66	1.12	NS	Medium	Large	NS
#Followers/#Followings	NS	0.25	0.36	NS	Small	Small
Nsfw	0.58	NS	0.92	Medium	NS	Large
Account's Age	1.08	0.30	0.70	Large	Small	Medium
#Tweets/Year	0.27	0.30	NS	Small	Small	NS
#Tweets	0.15	0.27	0.18	Very Small	Small	Very Small
Porn	0.25	0.42	0.80	Small	Small	Medium
Letter G in Screen Name	0.57	0.80	0.12	Medium	Large	Very Small
Organization	0.39	0.76	NS	Small	Medium	NS
#Followers	NS	0.32	0.33	NS	Small	Small
LGBT	0.52	0.71	0.10	Medium	Medium	Very Small
Username's Length	0.65	0.50	NS	Medium	Small	NS
Allies	0.34	0.65	NS	Small	Medium	NS
#Words in Bio	0.26	0.52	0.26	Small	Medium	Small
Events	0.34	0.64	NS	Small	Medium	NS
#Followings	NS	NS	0.13	NS	NS	Very Small
Men	0.10	0.68	0.45	Very Small	Medium	Small

4. Discussion

This research is unique in that it provides a prediction framework including an automatic classifier, a feature selection approach, and evaluation measures. Our experiments were designed to categorize LGBT users based on different sets of features and categories and identify features that may contribute to improving the efficiency and effectiveness of the prediction. Our proposed model uses BayesNet to learn feature vectors and the χ^2 value to identify the optimal subset of features. Our proposed model outperformed the baseline on classifying LGBT accounts. We are now able to identify individual, sex worker/porn, and organization accounts with around 88% accuracy. The evaluation shows that the performance of our classifier is better than the baseline accuracy (76.89%) using ZeroR, which classifies each user to the largest class, which is individual users in this study.

While even a little higher than the baseline could be significant, our classifier shows more than 10% improvement over the baseline.

While using profile and bio features independently can provide a significant change over the baseline performance, the combination of profile and bio features and reducing the number of features can be more helpful in classifying LGBT accounts. The accuracy of our classifier is improved when both profile and bio features are used. While the number of profile features (81) is less than the number of bio features (1288), most of the top 50 features are related to profile information, indicating that profile information containing structured features plays an important role in classifying LGBT accounts. In addition, words in the bio of Ind, SWP, and Org accounts can be a good indicator to categorize LGBT users.

Our results suggest that profile information, words of bio, and characters of username and screen name can help to predict the category of LGBT users. For example, it is not surprising to see that the number of followers of Org and SWP accounts is more than Ind accounts because they have more fans than Ind ones. However, it is interesting to find that Org accounts used the like icon and tweeted less than Ind and SWP accounts, which means that Org accounts are cautious in posting social comments and showing their interests. The reason behind this strategy could be that a single unfortunate post can have a significant negative impact on organizations [80]. However, Ind and SWP accounts do not have this limitation and can be more active than Org accounts.

Compared to SWP and Ind accounts, Org ones use community in the bio more than Ind and SWP accounts, which means Org accounts are more interested in emphasizing their role for the community. The age of Org accounts is higher than the other two accounts, which indicates that organizations have been active on social media for more years than the other two types. The characteristics of SWP accounts are similar to Org accounts based on some features. For example, the number of followers and followings of Org and SWP accounts is more than Ind accounts. Org and SWP accounts use more words in their bio than Ind accounts to introduce their services and provide more information for customers.

The comparisons of Ind vs. SWP, Ind vs. Org, and SWP vs. Org illustrate that the minimum difference is between Ind and SWP accounts, indicating that SWP accounts are more similar to Ind accounts than Org ones. It seems that the strategy of SWP accounts is to behave similarly to Ind accounts. Therefore, it is a complicated task to distinguish between Ind and SWP accounts. However, identifying Org accounts is less complicated than SWP and Ind accounts. While it is a difficult task to identify Ind and SWP accounts, there are features (e.g., using nsfw in bio) that assist in finding SWP accounts.

This research provides significant contributions. First, while other research developed binary classifiers, this paper offers a multi-label classifier to categorize users. For example, one study identifies individual and organization users [81]. Second, this study illustrates that the used traditional machine learning methods in this research offer better performance than deep learning using CNN for categorizing LGBT users using bio and profile features. Our data size is not very large. Therefore, this finding is in line with the current literature that indicates that deep learning methods do not provide a significant performance over traditional methods if the size of a dataset is small or medium [82]. Third, the proposed approach is effective in utilizing bio and profile features to identify Ind, SWP, and Org accounts. Fourth, this paper identifies and uses features that can be used for similar purposes. Fifth, the proposed approach is flexible to incorporate not only bio and profile features but also other features (e.g., semantics of tweets), use other machine learning methods, and be applied on other social media platforms. Sixth, this research is beneficial for researchers who are interested in categorized LGBT users for social media analysis purposes. For instance, our work can be used by public health experts to identify LGBT individuals to study their information behavior on social media, by social media and marketing companies and application developers to filter out adult content, and by social science and business experts to study LGBT organizations. We believe our work bears the potential to help understand the needs of LGBT individuals on social media and develop interventions to address the needs of LGBT people.

While our study contributes to LGBT studies in social media and opens a new direction for future research, this study bears certain limitations. First, we limit our features to bio and profile features. Second, this study is limited to LGBT users who live in the U.S. and post tweets in English. Third, our data collection was limited to lesbian, gay, bisexual, and transgender users, indicating that we might miss other possible relevant data.

Despite the limitations, our findings can provide new insights into types of LGBT users and their social media activities. Future research will need to consider n-grams (e.g., bigrams), linguistics features (e.g., verbs), the semantic meanings of words (e.g., themes), and global or local weighting methods. We aim to go beyond unigrams and incorporate n-grams, linguistic analysis, and semantic features in our prediction framework. That way, we hope to achieve a higher prediction level.

5. Conclusions

Twitter is a popular platform to obtain and analyze publicly available social media data. This platform has been used by researchers studying LGBT issues such as health. However, not all LGBT users are individual users. This research proposes a framework to categorize LGBT users on Twitter. We specially obtained features of Twitter accounts and developed an automated classifier with around 88% accuracy for categorizing LGBT users by type—user, sex worker/porn, and organizations. Our experiments were based on analyzing more than 16,000 Twitter accounts and showed that different types of LGBT accounts have distinct characteristics in their Twitter accounts, assisting in developing robust classifiers.

This research classifies LGBT users in three classes and explores several classification methods to identify the best classifier. Future work can address the limitations of this study, identify new features, develop classifiers with other machine learning techniques, and extend this work to other possible areas.

Author Contributions: Conceptualization, A.K.; methodology, A.K., F.W. and M.L.; software, A.K.; validation, A.K., M.L., F.W., H.R.B., M.Z. and D.L.; formal analysis, A.K., F.W., M.L., H.R.B., M.Z. and D.L.; investigation, A.K., M.L. and F.W.; resources, A.K.; data curation, A.K. and F.W.; writing—original draft preparation, A.K.; writing—review and editing, A.K., M.L., F.W. and H.R.B.; visualization, A.K.; supervision, A.K.; project administration, A.K.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Big Data Health Science Center (BDHSC) at the University of South Carolina. All opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gallup. LGBT Identification Rises to 5.6% in Latest, U.S. Estimate. 2021. Available online: https://news.gallup.com/poll/329708/lgbt-identification-rises-latest-estimate.aspx (accessed on 1 April 2021).
2. Gonzales, G.; Przedworski, J.; Henning-Smith, C. Comparison of health and health risk factors between lesbian, gay, and bi-sexual adults and heterosexual adults in the United States: Results from the National Health Interview Survey. *JAMA Intern. Med.* **2016**, *176*, 1344–1351. [CrossRef]
3. Byron, P.; Rasmussen, S.; Wright, T.D.; Lobo, R.; Robinson, K.H.; Paradise, B. 'You learn from each other': LGBTIQ Young People's Mental Health Help-seeking and the RAD Australia Online Directory. Available online: https://researchdirect.westernsydney.edu.au/islandora/object/uws:38815 (accessed on 1 April 2021).
4. Seidenberg, A.B.; Jo, C.L.; Ribisl, K.M.; Lee, J.G.L.; Butchting, F.O.; Kim, Y.; Emery, S.L. A National Study of Social Media, Television, Radio, and Internet Usage of Adults by Sexual Orientation and Smoking Status: Implications for Campaign Design. *Int. J. Environ. Res. Public Health* **2017**, *14*, 450. [CrossRef]
5. Pew Research Center. A Survey of LGBT Americans. In: Pew Research Center's Social & Demographic Trends Project [Inter-net]. 2013. Available online: https://www.pewresearch.org/social-trends/2013/06/13/a-survey-of-lgbt-americans/ (accessed on 15 April 2021).
6. Byron, P.; Albury, K.; Evers, C. It would be weird to have that on Facebook: Young people's use of social media and the risk of sharing sexual health information. *Reprod. Health Matters* **2013**, *21*, 35–44. [CrossRef]

7. Karami, A.; Lundy, M.; Webb, F.; Dwivedi, Y.K. Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access* **2020**, *8*, 67698–67717. [CrossRef]
8. Karami, A.; Kadari, R.; Panati, L.; Nooli, S.; Bheemreddy, H.; Bozorgi, P. Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? *ISPRS Int. J. Geo Inf.* **2021**, *10*, 373. [CrossRef]
9. Karami, A.; Dahl, A.; Shaw, G.; Valappil, S.; Turner-McGrievy, G.; Kharrazi, H.; Bozorgi, P. Analysis of Social Media Discussions on (#)Diet by Blue, Red, and Swing States in the U.S. *Healthcare* **2021**, *9*, 518. [CrossRef]
10. Karami, A.; Anderson, M. Social media and COVID-19, Characterizing anti-quarantine comments on Twitter. In Proceedings of the Association for Information Science and Technology, online, 22 October–1 November 2020; Volume 57, p. e349.
11. Karami, A.; Dahl, A.A.; Turner-McGrievy, G.; Kharrazi, H.; Shaw, G., Jr. Characterizing diabetes, diet, exercise, and obesity com-ments on Twitter. *Int. J. Inf. Manag.* **2018**, *38*, 1–6. [CrossRef]
12. Money, V.; Karami, A.; Turner-McGrievy, B.; Kharrazi, H. Seasonal characterization of diet discussions on Reddit. In Proceedings of the Proceedings of the Association for Information Science and Technology, online, 22 October–1 November 2020; Volume 57, p. 320.
13. Kordzadeh, N. Exploring the Use of Twitter by Leading Medical Centers in the United States. In Proceedings of the 52nd Hawaii International Conference on System Sciences, Grand Wailea, HI, USA, 8–11 January 2019.
14. Li, Z.; Qiao, S.; Jiang, Y.; Li, X. Building a Social Media-Based HIV Risk Behavior Index to Inform the Prediction of HIV New Diagnosis: A Feasibility Study. *AIDS* **2021**, *35*, S91–S99. [CrossRef]
15. Karami, A.; Elkouri, A. Political Popularity Analysis in Social Media. In Proceedings of the International Conference on Information (iConference), Washington, DC, USA, 31 March–3 April 2019.
16. Karami, A.; Bennett, L.S.; He, X. Mining public opinion about economic issues: Twitter and the us presidential election. *Int. J. Strateg. Decis. Sci.* **2018**, *9*, 18–28. [CrossRef]
17. Najafabadi, M.M.; Domanski, R.J. Hacktivism and distributed hashtag spoiling on Twitter: Tales of the #IranTalks. *First Monday* **2018**, *23*. [CrossRef]
18. Karami, A.; Spinel, M.; White, C.; Ford, K.; Swan, S. A Systematic Literature Review of Sexual Harassment Studies with Text Mining. *Sustainability* **2021**, *13*, 6589. [CrossRef]
19. Karami, A.; Shah, V.; Vaezi, R.; Bansal, A. Twitter speaks: A case of national disaster situational awareness. *J. Inf. Sci.* **2019**, *46*, 313–324. [CrossRef]
20. Turner-McGrievy, G.; Karami, A.; Monroe, C.; Brandt, H.M. Dietary pattern recognition on Twitter: A case example of before, during, and after four natural disasters. *Nat. Hazards* **2020**, *103*, 1035–1049. [CrossRef]
21. Martín, Y.; Cutter, S.L.; Li, Z. Bridging twitter and survey data for evacuation assessment of Hurricane Matthew and Hurri-cane Irma. *Nat. Hazards Rev.* **2020**, *21*, 04020003. [CrossRef]
22. Dzurick, A. *Lesbian, Gay, Bisexual, and Transgender Americans at Risk: Problems and Solutions*; Praeger: Santa Barbara, CA, USA, 2018; Social media, iPhones, iPads, and identity: Media impact on the coming-out process for LGBT youths.
23. Haimson, O.L.; Veinot, T.C. Coming Out to Doctors, Coming Out to "Everyone": Understanding the Average Sequence of Transgender Identity Disclosures Using Social Media Data. *Transgender Health* **2020**, *5*, 158–165. [CrossRef]
24. Khatua, A.; Cambria, E.; Ghosh, K.; Chaki, N.; Khatua, A. Tweeting in support of LGBT? A deep learning approach. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, Kolkata, India, 3–5 January 2019; pp. 342–345.
25. Selkie, E.; Adkins, V.; Masters, E.; Bajpai, A.; Shumer, D. Transgender Adolescents' Uses of Social Media for Social Support. *J. Adolesc. Health* **2020**, *66*, 275–280. [CrossRef]
26. Blackwell, L.; Hardy, J.; Ammari, T.; Veinot, T.; Lampe, C.; Schoenebeck, S. LGBT parents and social media: Advocacy, privacy, and disclosure during shifting social movements. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2016; pp. 610–622.
27. Guillory, J.; Wiant, K.F.; Farrelly, M.; Fiacco, L.; Alam, I.; Hoffman, L.; Crankshaw, E.; Delahanty, J.; Alexander, T.N. Recruiting hard-to-reach populations for survey re-search: Using Facebook and Instagram advertisements and in-person intercept in LGBT bars and nightclubs to recruit LGBT young adults. *J. Med. Internet Res.* **2018**, *20*, e197. [CrossRef]
28. Webb, F.; Karami, A.; Kitzie, V.L. Characterizing Diseases and Disorders in Gay Users' Tweets. In Proceedings of the Southern Association for Information Systems (SAIS), Atlanta, GA, USA, 23 March 2018.
29. Karami, A.; Webb, F.; Kitzie, V.L. Characterizing transgender health issues in Twitter. In Proceedings of the Association for Information Science and Technology, Vancouver, BC, Canada, 4–9 November 2018; Volume 55, pp. 207–215.
30. Karami, A.; Webb, F. Analyzing health tweets of LGB and transgender individuals. In Proceedings of the Association for Information Science and Technology, online, 22 October–1 November 2020; Volume 57, p. 264.
31. Carrasco, M.; Kerne, A. Queer visibility: Supporting LGBT+ selective visibility on social media. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–12.
32. Escobar-Viera, C.G.; Whitfield, D.L.; Wessel, C.B.; Shensa, A.; E Sidani, J.; Brown, A.L.; Chandler, C.J.; Hoffman, B.L.; Marshal, M.P.; A Primack, B. For better or for worse? A systematic re-view of the evidence on social media use and depression among lesbian, gay, and bisexual minorities. *JMIR Ment. Health* **2018**, *5*, e10496. [CrossRef]
33. Han, X.; Han, W.; Qu, J.; Li, B.; Zhu, Q. What happens online stays online?—Social media dependency, online support behavior and offline effects for LGBT. *Comput. Hum. Behav.* **2019**, *93*, 91–98. [CrossRef]

34. Hswen, Y.; Sewalk, K.C.; Alsentzer, E.; Tuli, G.; Brownstein, J.S.; Hawkins, J.B. Investigating inequities in hospital care among lesbian, gay, bisexual, and transgender (LGBT) individuals using social media. *Soc. Sci. Med.* **2018**, *215*, 92–97. [CrossRef]
35. Haimson, O. Mapping gender transition sentiment patterns via social media data: Toward decreasing transgender mental health disparities. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 749–758. [CrossRef] [PubMed]
36. Krueger, E.A.; Young, S.D. Twitter: A Novel Tool for Studying the Health and Social Needs of Transgender Communities. *JMIR Ment. Health* **2015**, *2*, e16. [CrossRef]
37. Gold, J.; Pedrana, A.; Stoove, M.; Chang, S.; Howard, S.; Asselin, J.; Ilic, O.; Batrouney, C.; Hellard, M.E. Developing Health Promotion Interventions on Social Networking Sites: Recommendations from The FaceSpace Project. *J. Med. Internet Res.* **2012**, *14*, e30. [CrossRef] [PubMed]
38. Pedrana, A.; Hellard, M.; Gold, J.; Ata, N.; Chang, S.; Howard, S.; Asselin, J.; Ilic, O.; Batrouney, C.; Stoove, M.; et al. Queer as F**k: Reaching and Engaging Gay Men in Sexual Health Promotion through Social Networking Sites. *J. Med. Internet Res.* **2013**, *15*, e25. [CrossRef] [PubMed]
39. McDaid, L.M.; Lorimer, K. P5.044 A Proactive Approach to Online Chlamydia Screening: Qualitative Exploration of Young Men's Perspectives of the Barriers and Facilitators. *Sex. Transm. Infect.* **2013**, *89*, A348. [CrossRef]
40. Wohlfeiler, D.; Hecht, J.; Volk, J.; Raymond, H.F.; Kennedy, T.; McFarland, W. How can we improve online HIV and STD preven-tion for men who have sex with men? Perspectives of hook-up website owners, website users, and HIV/STD directors. *AIDS Behav.* **2013**, *17*, 3024–3033. [CrossRef] [PubMed]
41. Young, S.D.; Harrell, L.; Jaganath, D.; Cohen, A.C.; Shoptaw, S. Feasibility of recruiting peer educators for an online social networking-based health intervention. *Health Educ. J.* **2013**, *72*, 276–282. [CrossRef]
42. 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]
43. Mustanski, B.; Greene, G.J.; Ryan, D.; Whitton, S.W. Feasibility, Acceptability, and Initial Efficacy of an Online Sexual Health Promotion Program for LGBT Youth: The Queer Sex Ed Intervention. *J. Sex Res.* **2014**, *52*, 220–230. [CrossRef]
44. Gabarron, E.; Wynn, R. Use of social media for sexual health promotion: A scoping review. *Glob. Health Action* **2016**, *9*, 32193. [CrossRef]
45. Martinez, O.; Wu, E.; Shultz, A.Z.; Capote, J.; Rios, J.L.; Sandfort, T.; Manusov, J.; Ovejero, H.; Carballo-Dieguez, A.; Baray, S.C.; et al. Still a Hard-to-Reach Population? Using Social Media to Recruit Latino Gay Couples for an HIV Intervention Adaptation Study. *J. Med. Internet Res.* **2014**, *16*, e113. [CrossRef] [PubMed]
46. Elliot, E.; Rossi, M.; McCormack, S.; McOwan, A. Identifying undiagnosed HIV in men who have sex with men (MSM) by offering HIV home sampling via online gay social media: A service evaluation. *Sex. Transm. Infect.* **2016**, *92*, 470–473. [CrossRef]
47. Rhodes, S.D.; McCoy, T.P.; Tanner, A.E.; Stowers, J.; Bachmann, L.H.; Nguyen, A.L.; Ross, M. Using Social Media to Increase HIV Testing Among Gay and Bisexual Men, Other Men Who Have Sex With Men, and Transgender Persons: Outcomes From a Randomized Community Trial. *Clin. Infect. Dis.* **2016**, *62*, 1450–1453. [CrossRef] [PubMed]
48. Reiter, P.L.; Katz, M.L.; A Bauermeister, J.; Shoben, A.B.; Paskett, E.D.; McRee, A.-L. Recruiting Young Gay and Bisexual Men for a Human Papillomavirus Vaccination Intervention through Social Media: The Effects of Advertisement Content. *JMIR Public Health Surveill.* **2017**, *3*, e33. [CrossRef]
49. Cao, B.; Liu, C.; Durvasula, M.; Tang, W.; Pan, S.; Saffer, A.J.; Wei, C.; Tucker, J.D.; Jiang, J.; Zhao, H. Social Media Engagement and HIV Testing Among Men Who Have Sex With Men in China: A Nationwide Cross-Sectional Survey. *J. Med. Internet Res.* **2017**, *19*, e251. [CrossRef]
50. Patel, V.V.; Ginsburg, Z.; A Golub, S.; Horvath, K.J.; Rios, N.; Mayer, K.H.; Kim, R.S.; Arnsten, J.H. Empowering With PrEP (E-PrEP), a Peer-Led Social Media–Based Intervention to Facilitate HIV Preexposure Prophylaxis Adoption among Young Black and Latinx Gay and Bisexual Men: Protocol for a Cluster Randomized Controlled Trial. *JMIR Res. Protoc.* **2018**, *7*, e11375. [CrossRef] [PubMed]
51. Qureshi, R.; Zha, P.; Kim, S.; Hindin, P.; Naqvi, Z.; Holly, C.; Dubbs, W.; Ritch, W. Health Care Needs and Care Utilization Among Lesbian, Gay, Bisexual, and Transgender Populations in New Jersey. *J. Homosex.* **2018**, *65*, 167–180. [CrossRef]
52. Tanner, A.E.; Song, E.Y.; Mann-Jackson, L.; Alonzo, J.; Schafer, K.; Ware, S.; Garcia, J.M.; Hall, E.A.; Bell, J.C.; Van Dam, C.N.; et al. Preliminary Impact of the weCare Social Media Intervention to Support Health for Young Men Who Have Sex with Men and Transgender Women with HIV. *Aids Patient Care STDs* **2018**, *32*, 450–458. [CrossRef]
53. Card, K.G.; Lachowsky, N.; Hawkins, B.W.; Jollimore, J.; Baharuddin, F.; Hogg, R.S.; Willoughby, J.; Bauermeister, J.; Zlotorzynska, M.; Kite, J. Predictors of Facebook User Engagement with Health-Related Content for Gay, Bisexual, and Other Men Who Have Sex With Men: Content Analysis. *JMIR Public Health Surveill.* **2018**, *4*, e38. [CrossRef]
54. Verrelli, S.; White, F.; Harvey, L.; Pulciani, M.R. Minority stress, social support, and the mental health of lesbian, gay, and bisexual Australians during the Australian Marriage Law Postal Survey. *Aust. Psychol.* **2019**, *54*, 336–346. [CrossRef]
55. Kruger, S.; Hermann, B. Can an Online Service Predict Gender? On the State-of-the-Art in Gender Identification from Texts. In Proceedings of the 2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE), Montreal, QC, Canada, 27–29 May 2019.

56. Rangel, F.; Rosso, P.; Montes-y-Gómez, M.; Potthast, M.; Stein, B. Overview of the 6th Author Profiling Task at Pan 2018, Multi-Modal Gender Identification in Twitter. Available online: http://personales.upv.es/prosso/resources/RangelEtAl_PAN18.pdf (accessed on 15 April 2021).
57. Burger, J.D.; Henderson, J.; Kim, G.; Zarrella, G. Discriminating Gender on Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1301–1309.
58. Wang, Q.; Ma, S.; Zhang, C. Predicting users' demographic characteristics in a Chinese social media network. *Electron. Libr.* **2017**, *35*, 758–769. [CrossRef]
59. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.; Dziurzynski, L.; Ramones, S.M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M.E.P.; et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* **2013**, *8*, e73791. [CrossRef]
60. Peersman, C.; Daelemans, W.; Van Vaerenbergh, L. Predicting age and gender in online social networks. In Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, Glasgow, UK, 24–28 October 2011; pp. 37–44.
61. Yang, K.-C.; Varol, O.; Hui, P.-M.; Menczer, F. Scalable and Generalizable Social Bot Detection through Data Selection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 1096–1103.
62. Kamath, C.N.; Bukhari, S.S.; Dengel, A. Comparative study between traditional machine learning and deep learning approach-es for text classification. In Proceedings of the ACM Symposium on Document Engineering, Halifax, NS, Canada, 28–31 August 2018; pp. 1–11.
63. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [CrossRef]
64. Zhang, C.; Liu, C.; Zhang, X.; Almpanidis, G. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Syst. Appl.* **2017**, *82*, 128–150. [CrossRef]
65. Fernández-Delgado, M.; Cernadas, E.; Barro, S.; Amorim, D. Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.* **2014**, *15*, 3133–3181.
66. Pham, B.T.; Bui, D.T.; Prakash, I. Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study. *Geotech. Geol. Eng.* **2017**, *35*, 2597–2611. [CrossRef]
67. Chimieski, B.F.; Fagundes, R.D.R. Association and classification data mining algorithms comparison over medical datasets. *J. Health Inform.* **2013**, *5*, 44–51.
68. Zhao, Y.; Zhang, Y. Comparison of decision tree methods for finding active objects. *Adv. Space Res.* **2008**, *41*, 1955–1959. [CrossRef]
69. Bassem, B.; Zrigui, M. Gender Identification: A Comparative Study of Deep Learning Architectures. In Proceedings of the Advances in Intelligent Systems and Computing, Vellore, India, 6–8 December 2019.
70. Sezerer, E.; Polatbilek, O.; Sevgili, Ö.; Tekir, S. Gender prediction from Tweets with convolutional neural networks: Notebook for PAN at CLEF 2018. In Proceedings of the 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF CEUR Workshop Proceedings, Avignon, France, 10–14 September 2018.
71. Wei, F.; Qin, H.; Ye, S.; Zhao, H. Empirical Study of Deep Learning for Text Classification in Legal Document Review. In Proceedings of the 2018 IEEE International Conference on Big Data, Seattle, WA, USA, 10–13 December 2018; pp. 3317–3320.
72. Karami, A.; Zhou, B. Online Review Spam Detection by New Linguistic Features. In Proceedings of the iConference, Irvine, CA, USA, 24–27 March 2015.
73. Karami, A.; Zhou, L. Exploiting latent content based features for the detection of static SMS spams. *Proc. Am. Soc. Inf. Sci. Technol.* **2014**, *51*, 1–4. [CrossRef]
74. Karami, A.; Swan, S.; Moraes, M.F. Space identification of sexual harassment reports with text mining. In Proceedings of the Association for Information Science and Technology, online, 22 October–1 November 2020; Volume 57, p. 265.
75. Yang, Y.; Pedersen, J.O. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 1997; pp. 412–420.
76. Tukey, J.W. Comparing Individual Means in the Analysis of Variance. *Biometrics* **1949**, *5*, 99. [CrossRef]
77. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
78. Jafari, M.; Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell J.* **2019**, *20*, 604–607. [PubMed]
79. Sawilowsky, S.S.; Salkind, N. Journal of Modern Applied Statistical Methods. *Encycl. Meas. Stat.* **2013**, *26*. [CrossRef]
80. Ollier-Malaterre, A.; Rothbard, N.P. How to Separate the Personal and Professional on Social Media. Harvard Business Re-view. Available online: https://hbr.org/2015/03/how-to-separate-the-personal-and-professional-on-social-media (accessed on 23 July 2021).
81. Wood-Doughty, Z.; Mahajan, P.; Dredze, M. Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 56–61. [CrossRef]
82. Zhao, W. Research on the deep learning of the small sample data based on transfer learning. In Proceedings of the AIP Conference Proceedings, Bydgoszcz, Poland, 9–11 May 2017; p. 020018.