Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Finding Sensitive Accounts on Twitter: An Automated Approach Based on Follower Anonymity

Sai Teja Peddinti[⋆] and **Keith W. Ross**[⋆ †] and **Justin Cappos**[⋆]
{psaiteja, keithwross, jcappos}@nyu.edu
[⋆]Dept. of Computer Science and Engineering, New York University
[†]New York University Shanghai

Abstract

We explore the feasibility of *automatically* finding accounts that publish sensitive content on Twitter, by examining the percentage of anonymous and identifiable followers the accounts have. We first designed a machine learning classifier to automatically determine if a Twitter account is anonymous or identifiable. We then classified an account as potentially sensitive based on the percentages of anonymous and identifiable followers the account has. We applied our approach to approximately 100,000 accounts with 404 million active followers. The approach uncovered accounts that were sensitive for a diverse number of reasons.

Introducing appropriate privacy controls across online services necessitates defining and identifying sensitive content, that is, content that needs special consideration and protection. However, there isn't a single definition shared by the legal and data protection authorities and the online service providers about what constitutes sensitive content. This lack of a universal definition makes it hard to simply enumerate a list of sensitive content categories.

In this paper we consider identifying sensitive content on Twitter. Specifically, we seek to develop an efficient automated means for identifying accounts that tweet sensitive content. Such an automated means for identifying sensitive accounts can shed significant insight on contemporary social media, and aid in updating privacy features/controls and policies.

Our Approach

Twitter does not enforce a real-name policy, enabling some users to adopt non-identifying pseudonyms (termed *anonymous* accounts) and others to voluntarily reveal their identities by disclosing their full names (termed *identifiable* accounts). A recent study analyzed accounts relating to sensitive topics (such as pornography, religious hatred, and drugs) and non-sensitive topics (such as news and family), and found that the sensitive accounts have relatively large percentages of anonymous followers and relatively small percentages of identifiable followers, and vice versa for the non-sensitive accounts (Peddinti, Ross, and Cappos 2014).

In this work, we label a Twitter account to be potentially sensitive if it has a relatively large number of anonymous followers and a relatively small number of identifiable followers. To automatically find these sensitive accounts on Twitter, we first consider the sub-problem of automatically determining if a Twitter account is anonymous or identifiable. We then develop a heuristic for classifying an account as sensitive as a function of the percentages of anonymous and identifiable followers that the account has. We applied our approach to approximately 100,000 accounts with 404 million active followers. In addition to detecting many of the usual suspects (accounts related to pornography, drugs, and so on), our approach uncovered many accounts related to socially stigmatized topics, such as depression, self-mutilation, obesity, and anorexia.

Background

For our research we primarily use the information present in the profile of a Twitter account. The profile includes a unique alphanumeric ID (sometimes called the screen name), a name string, a description, a profile picture, location information, and a URL field (to link other websites). Each Twitter account has friends (accounts it follows) and followers (accounts that follow it), but we emphasize more on followers in our study.

Similar to prior work (Peddinti, Ross, and Cappos 2014), we categorize Twitter users based on their degree of anonymity:

- **Anonymous** – A Twitter account containing neither the first nor last name, and not containing a URL in the profile (which may point to a web page that identifies or partially identifies the user).
- **Partially Anonymous** – A Twitter account having a first name or a last name but not both in the profile.
- **Identifiable** – A Twitter account containing both a first name and a last name in the profile.
- **Unclassifiable** – A Twitter account that is neither Anonymous, Identifiable, nor Partially Anonymous. Accounts which have neither a first nor a last name but have a URL fall under this category.

Twitter is plagued by unused ephemeral accounts or those created to spread spam (Grier et al. 2010). To avoid any bias

on the results caused by these accounts, we remove from our data sets all accounts that have no friends and followers or haven't posted a tweet six months after account creation (ephemeral). We eliminated accounts that have some resemblance to reported spam account behavior. In addition, we also sanitized our datasets by eliminating all accounts which do not report English as the language of preference.

Twitter Data Sets

Labeled Training Data

We used supervised machine learning to automatically classify accounts as sensitive, and also classify the account followers as anonymous or identifiable. For this we leveraged labeled data from the prior study (Peddinti, Ross, and Cappos 2014), which contains two distinct data sets. The first data set measures the prevalence of anonymity on Twitter using a random accounts sample. The second dataset studies the influence of content sensitivity on user anonymity, and it was created by picking 47 Twitter accounts related to different sensitive categories (such as pornography and escort services), and 20 accounts related to non-sensitive categories (such as news sites and family recreation). The followers of these 67 accounts were sanitized (removed ephemeral, spam, and non-English accounts) and labeled. The combined labeled accounts from the two data sets constitute our training set. The distribution of the accounts across the different anonymity categories is shown in Table 1.

Label	# of Twitter Accounts
Identifiable	66,903 (51.3%)
Partially Anonymous	27,734 (21.2%)
Anonymous	19,890 (15.2%)
Unclassifiable	16,105 (12.3%)
Total	130,632

Table 1: Training Set for Machine Learning

Crawled Test Data Set

To validate if our approach can identify potentially sensitive Twitter accounts, we created a new test dataset by crawling Twitter from May 31 - Aug 7, 2014. Starting from the 67 hand-picked accounts in the training set belonging to the sensitive and non-sensitive topics (the seed list), we crawled outwards and collected more than 100,000 accounts with a total of half billion followers. We sanitized the resulting set by removing all non-English accounts and accounts with $<$200 *active* (non-ephemeral and non-spam) followers. Our resulting data set has 93,042 accounts with approximately 404 million active followers. We applied our sensitive account discovery methodology to this data set.

Automating Identification of Anonymous Accounts

Since the definitions for all the user groups rely on the presence/absence of first/last names in the Twitter account's profile, we obtained public name lists from the United States Census [1] and Social Security Administration [2] databases. Checking for memberships in these name lists yielded very poor anonymous and identifiable detection rates, and one of the primary reasons was the occurrence of common English words in the name lists.

A Twitter profile has other properties in addition to names. We utilized all these properties as input to a machine learning classifier. In addition to the presence or absence of first/last names, we considered if the name string was structured (such as *FirstName MiddleName LastName*, *FirstName MiddleInitial LastName*, or *FirstName LastName*), and take into account the popularity ranks of the occurring names in the name lists. To limit classification errors due to English words occurring in name lists, we leveraged word dictionaries used in the Scrabble board game (as they generally do not contain proper nouns)[3] and word frequencies obtained from the British National Corpus[4].

We considered the number of friends, followers, tweets, and *favorited* tweets an account has. Twitter users can be grouped into *lists* for easy reading of tweets. We considered the number of lists an account has membership in. We checked if a Twitter profile hides its activity using the *protected* feature and shares its location using the *geo-tagging* feature. We also considered the profile pictures, but they weren't very helpful in deducing the identity of an individual based on our initial exploration, so we did not include them in our study. After testing various configurations with the features and feature representations, we chose 16 features: 12 numeric and 4 boolean, and they are listed in Table 2.

Type	Feature
Numeric	# of friends
	# of followers
	followers-to-friends ratio
	# of user list memberships
	# of tweets
	# of favorite tweets
	number of parts/words in the name string
	popularity rank of occurring first name
	popularity rank of occurring last name
	# of Scrabble words present in the name
	word frequency rank of occurring first name in the Scrabble list
	word frequency rank of occurring last name in Scrabble list
Boolean	enabled *protected* privacy feature
	enabled *geo-tagging* for tweets
	includes a url in the profile
	name follows structural constraints

Table 2: Selected Feature Set for Machine Learning Classification

[1] http://www.census.gov/genealogy/www/data/index.html

[2] http://www.ssa.gov/oact/babynames/limits.html

[3] http://www.freescrabbledictionary.com/sowpods.txt and http://www.isc.ro/en/commands/lists.html

[4] http://www.kilgarriff.co.uk/bnc-readme.html

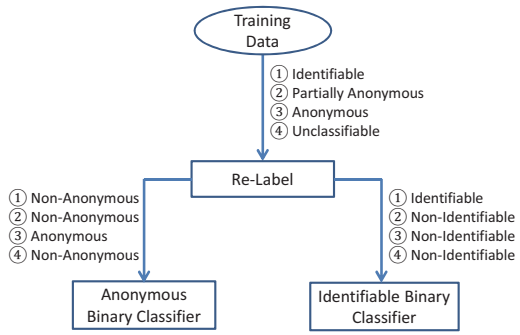

Figure 1: Machine Learning Training

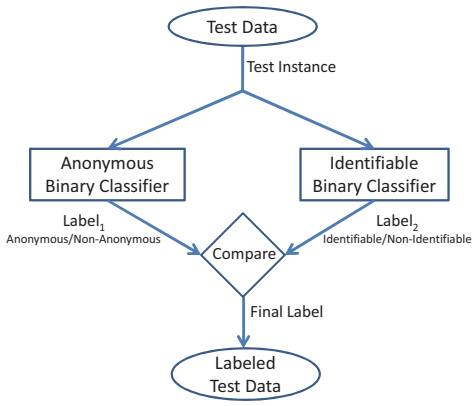

Figure 2: Machine Learning Testing

Customized Classifier for Account Anonymity Classification

Our dataset has four classes (Table 1). We developed a classifier that converts this four-class classification problem into two binary classification problems: one that classifies each account either as anonymous or non-anonymous; the other that classifies each account as either identifiable or non-identifiable. The results of the two classifiers are then combined to classify each account as "anonymous," "identifiable" or "unknown" as described below.

The training phase is shown in Figure 1. The training data containing four classes gets relabeled into two data sets containing the same number of training instances as the original. In the first data set, all the instances for classes other than the anonymous class get re-labeled as 'Non-Anonymous,' and this data set is passed to a binary classifier optimized for detecting anonymous accounts. In the second data set, all the instances for classes other than the identifiable get re-labeled as 'Non-Identifiable,' and this data set is passed to a binary classifier optimized for detecting identifiable accounts. Both the binary classifiers use Random Forest with 100 trees.

The testing phase is shown in Figure 2. Each test instance gets passed to each of the binary classifiers, which independently assigns a label to the instance. We determine the final label based on the decision table in Table 3. "Unknown" means we do not attempt to classify the account. Our account anonymity classifier had a precision of 0.9 and recall of 0.244 for anonymous accounts, and a precision of 0.932 and recall of 0.747 for Identifiable acccounts.

Label$_1$	**Label$_2$**	**Final Label**
Anonymous	Non-Identifiable	Anonymous
Non-Anonymous	Identifiable	Identifiable
Non-Anonymous	Non-Identifiable	Unknown
Anonymous	Identifiable	Unknown *(Did not occur)*

Table 3: Deciding Final Label for a Test Instance

Sensitive Account Discovery

The accounts identified in the previous section are referred to as "discovered anonymous" and "discovered identifiable" accounts, and we use these as proxies for the actual anonymous and identifiable accounts to find sensitive accounts. As mentioned earlier, we suspect that an arbitrary account is potentially sensitive (non-sensitive) if it has a relatively large (small) number of anonymous followers and a relatively small (large) number of identifiable users. To identify what large and small percentage values for discovered anonymous and identifiable accounts make sense, we classified the followers of the 67 accounts in the training dataset and determined the fractions of discovered anonymous and identifiable followers for each account. Figure 3 shows a scatter diagram, where each circle (triangle) corresponds to one of the chosen sensitive (non-sensitive) accounts. Strikingly, the sensitive accounts all lie at the top-left, and the non-sensitive accounts all lie at the bottom-right of the plot. Using a Support Vector Machine (SVM) classifier, we can separate the sensitive and non-sensitive accounts. The linear hyperplane equation obtained is $y = 0.0575x + 0.0078$.

We say that we *suspect a Twitter account to be sensitive* if $y > 0.0575x + 0.0078$, where y is the fraction of discovered anonymous followers and x is the fraction of discovered identifiable followers for the account. Further, if $y >> 0.0575x + 0.0078$, we suspect the account to be **very sensitive**. In a similar manner, we suspect accounts to be non-sensitive and very non-sensitive by reversing the inequalities. For a given account, the x and y values are determined by the automatic classification technique described earlier.

When we applied this methodology to the 93,042 random test accounts, 59.3% of the accounts lie on the sensitive side of the linear hyperplane, and 40.7% lie on the non-sensitive side. Below we study the accounts that are on the sensitive side and are far away from the linear hyperplane, i.e., the very sensitive accounts.

Types of Very Sensitive Accounts

We manually inspected the top 300 very sensitive accounts identified by our methodology, and assigned each account to a theme. Table 4 lists these themes. The miscellaneous theme contains accounts that are of individuals (identifying as females), or ones that share multimedia, post non-English

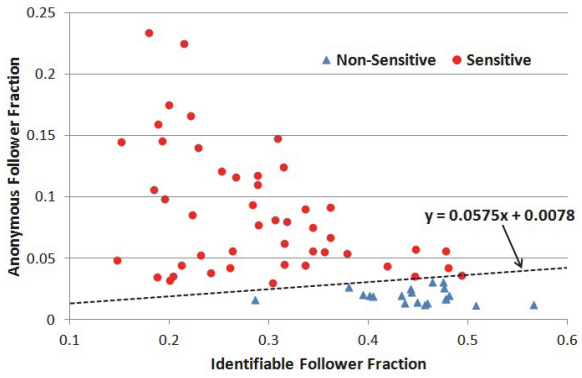

Figure 3: Scatter Plot of Sensitive and Non-Sensitive Accounts Based on Discovered Anonymous and Identifiable Followers

# of Sensitive Accounts	Theme
106	Couples sharing their intimate pictures or inviting swingers.
47	Arabic Adult/Gay Content
21	Relate to pornography/adult content.
12	Related to drugs, such as marijuana.
9	Accounts self-identifying as high school and college girls (some in their teens).
8	People obsessed with weight loss or anorexia.
6	Expressing depression, suicidal tendencies, or social anxieties.
5	Relate to Gay/Lesbian pornography.
5	Self-identifying as Female fitness/yoga accounts.
1	Groups supporting Lesbian, Gay, Bisexual, Transgender and Queer (LGBTQ).
80	Miscellaneous

Table 4: Top Sensitive Accounts and their Themes

tweets, or have adopted protected status or been de-activated after our initial data gathering.

As expected, pornography, drugs, and adult content are pervasive in the sensitive accounts. We identified several accounts discussing drugs, such as marijuana. There are accounts supporting and fighting for rights of lesbian, gay, bisexual and transgenders; accounts self-identifying as high school and college females; female fitness/yoga accounts; accounts that deal with severe cases of anorexia, social anxiety, depression and suicidal tendencies.

This preliminary examination clearly shows that the vast majority of automatically-identified potentially sensitive accounts are tweeting about topics that most people consider to be sensitive. These results also indicate that our methodology is generalizable – finds many sensitive topics (such as obesity and anorexia) which were not originally hand-chosen in the prior study (Peddinti, Ross, and Cappos 2014).

Related Work

There has not been much work on exploring the diversity of topics that online users consider sensitive. Researchers have tried to capture user content sensitivity preferences on a pre-determined list of topic categories (Rainie et al. ; Hawkey and Inkpen 2006), or relied on user self-reporting during surveys and interviews (Wang et al. 2011). These methodologies have limitations of being subjective, or are expensive. Prior research closest to our work is (Peddinti et al. 2014). It deals with identifying sensitive topic categories, while we focus on identifying sensitive user accounts. Unlike the earlier study, we do not rely on predefined content category tags, and are able to generalize to include overlooked topics.

Conclusion

We developed a novel and objective methodology, based on follower anonymity patterns, for identifying potentially sensitive accounts on Twitter. We applied it on a large Twitter crawl containing approximately 100,000 accounts with 404 million active followers, and uncovered sensitive accounts across diverse themes.

Acknowledgments

This work was supported in part by the NSF (under grant CNS-1318659). The views and conclusions contained in this document are those of the authors alone.

References

Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2010. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)*.

Hawkey, K., and Inkpen, K. M. 2006. Examining the content and privacy of web browsing incidental information. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*.

Peddinti, S. T.; Korolova, A.; Bursztein, E.; and Sampemane, G. 2014. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Proceedings of the 35th IEEE Symposium on Security & Privacy*.

Peddinti, S. T.; Ross, K. W.; and Cappos, J. 2014. "On the Internet, nobody knows you're a dog": A Twitter Case Study of Anonymity in Social Networks. In *Proceedings of the ACM Conference on Online Social Networks (COSN)*.

Rainie, L.; Kiesler, S.; Kang, R.; and Madden, M. Anonymity, Privacy, And Security Online: Part 4: How Users Feel About the Sensitivity of Certain Kinds of Data. http://www.pewinternet.org/2013/09/05/part-4-how-users-feel-about-the-sensitivity-of-certain- kinds-of-data/.

Wang, Y.; Norcie, G.; Komanduri, S.; Acquisti, A.; Leon, P. G.; and Cranor, L. F. 2011. 'I regretted the minute I pressed share': a qualitative study of regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS)*.