Automated Detection of Doxing on Twitter

Younes Karimi, Anna Squicciarini, Shomir Wilson

The 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing

November 8–22, 2022

Image : komando.com/privacy/people-use-doxing-to-get-revenge-online-protect-yourself/560317

PennState

Outline

PennState

Cyberbullying

Image : cheapsslshop.com/blog/cyber-bullying-facts-and-tips-to-prevent-it

What is Cyberbullying ?

- A subset of bullying
- Any type of online harassment performed using communication technologies by a user/group of users against other users
- Compared to traditional bullying :
  - Victims are more reachable for cyberbullying regardless of their physical locations
  - This can be more persistent

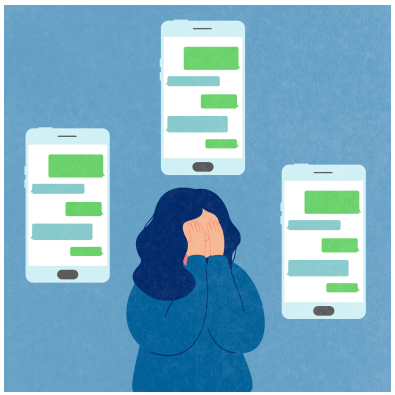

Image : parents.com/kids/problems/bullying

Image : magdalene.co/story/cyberbullied-in-indonesia

PennState

Background

What is Doxing ?

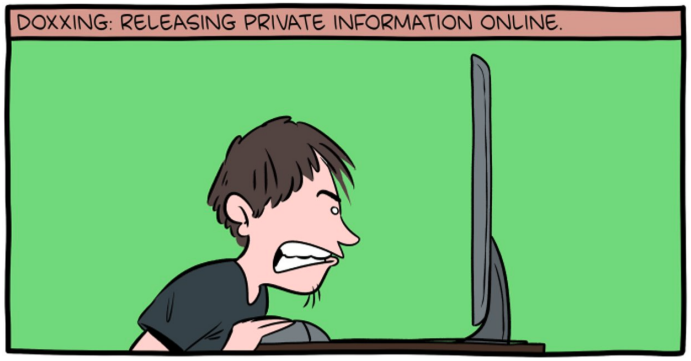

Image : smbc-comics.com/comic/dox

Image : nathatshjalpen.se/a/facts-doxing

- A type of cyberbullying
- Dox : An abbreviation for “documents”
- The act of disclosing sensitive information about others without their consent
- An unpleasant and sometimes dangerous phenomenon in microblogs and social media websites such as Twitter
- May put people’s careers or lives at risk (e.g., by encouraging kidnapping, child trafficking, intimidating) : Uninvolved man was shot dead in 2017 Wichita swatting
- Platforms : Instant Messenger, Social networking websites, Chatroom, Email, Video-sharing website, Webpage, Forum, Blog
- Attackers may be even known and close to the victims (e.g., parents, family members, classmates, friends outside the school, strangers, etc)

PennState

Background

Examples of Doxed Information

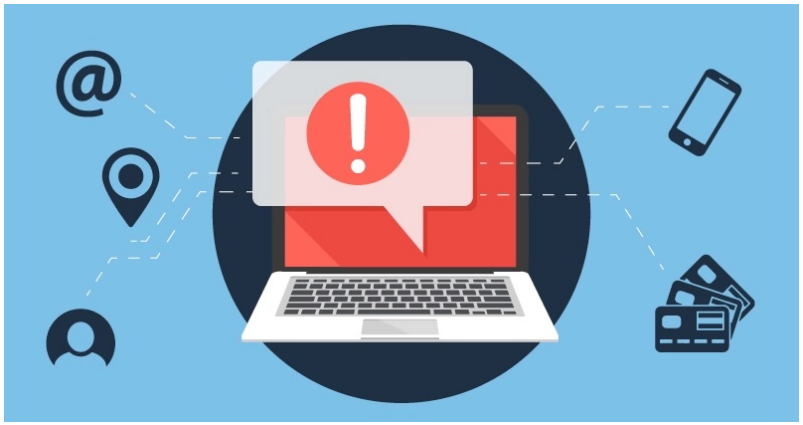

Image : vpnpro.com/wp-content/uploads/Doxxing-800x420-1.jpg

1. Demographic information : birthday, sexual orientation, race, ethnicity, and religion
2. Location info : street address, ZIP code, IP address, and GPS coordinates
3. Identity documents : passport and social security number
4. Contact info : phone number, email address
5. Financial information : credit card and bank account details
6. Sign-in credentials : usernames and passwords

PennState

Background

Information may be obtained through :

- Collecting and aggregating hard-to-access information
- Outing : Reshare
- Trickery : Social engineering, phishing, spear-phishing, impersonation

Image : protectimus.com/blog/doxing

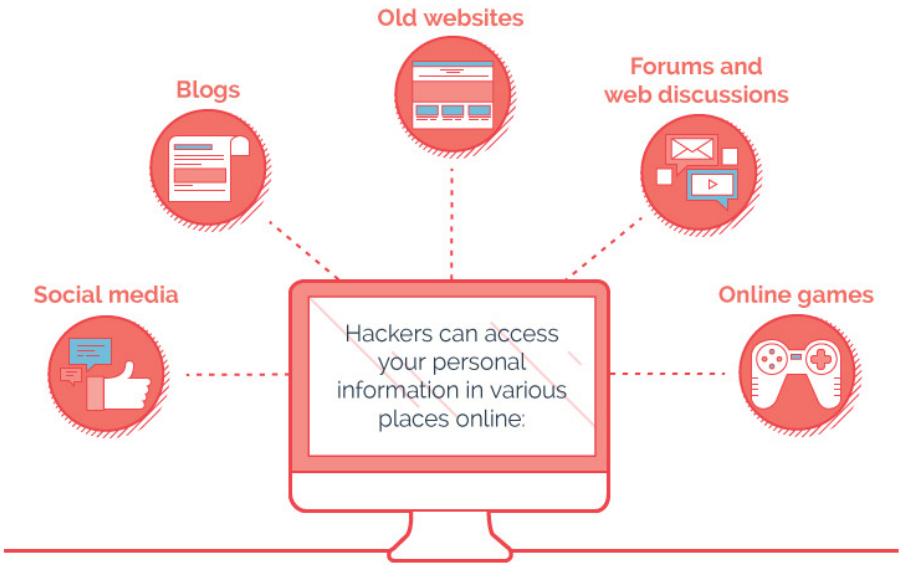

Image : diggitmagazine.com/articles/doxing-and-online-communities

PennState

Motivation

1. Literature has looked into detection of cyberbullying in general (disregarding private information), and privacy leakage (disregarding the actor)
2. No automated detection of doxing on Twitter that differentiates self-disclosures and non-sensitive disclosures from doxing and malicious disclosures
3. Once the private information is exposed to the public, it cannot be withdrawn
4. No comprehensive categorization of potential doxing intentions

Image : tumblr.com/search/doxxing

PennState

Comparison to the Literature

Method/Features	Detection Capability	NS/1st vs 2nd/3rd Party
Syntactic Heuristics [1]	Sensitive *location* and *emotion* information disclosure	Does not differentiate
Topics + Grammar Dependency [2]	Private information	Does not differentiate
Contextualized String Embedding (this work)	Doxing and malicious sensitive information disclosure	Differentiates disclosures

Comparison of similar works, their methods and detection capabilities. “NS” stands for “Nonsensitive” disclosures. None of the earlier works is capable of differentiating nonsensitive and self-disclosures (1st party) from doxing and malicious disclosures of second- and third-parties’ sensitive information.

Image : theatlantic.com/technology/archive/2014/03/doxing-an-etymology/284283

PennState

Summary of Contributions

1. Create and curate a dataset of 179,350 tweets that are likely to contain doxing episodes, and manually label 3,131 tweets.
2. Present a comprehensive insight into potential intentions behind sensitive information disclosures and differentiate between these motivations in terms of whether they constitute doxing or malicious disclosures and if they are defensive acts.
3. Describe an automated approach for detection of doxing and malicious sensitive information disclosures about second- and third-parties.

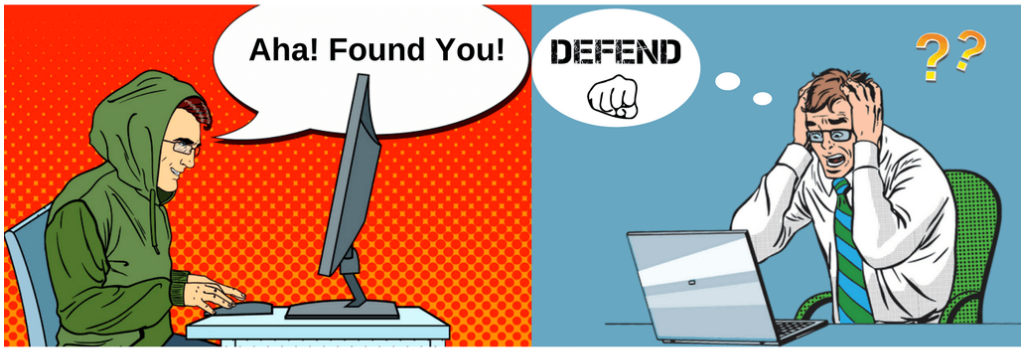

Image : limevpn.com/strengthen-your-defence-against-cyber-stalkers-and-doxing

PennState

Dataset

- Twitter public streaming API
- Collecting streams of tweets in real time
- Keyword-based collecting of tweets : An initial filter for discarding tweets that are less-likely to contain SSN and IP address
- Excludes tweets that are written in other languages

Dataset

Category	Sub-Category	Keywords	Tweets	Filtered & Labeled	$2^{nd}/3^{rd}$ Party	SD/NS
Identity Docs	SSN	ssn, ssa, social security number, social security administration	140,510	520	219	301
Location Info	IP Address	ip address	38,840	2,611	1916	695

Sensitive information categories, corresponding keywords and collected tweets. “SD” and “NS” stand for “Self-Disclosure” and “Nonsensitive” respectively. Additionally, “$2^{nd}/3^{rd}$ Party” tweets are those that disclose sensitive information about others.

Dataset

- Ground truth
- Manual inspection by human experts (based on pre-defined criteria)
- Labeling assessment
- A sample of 200 tweets labeled by second and third annotators
- Kappa and Fleiss' Kappa coefficients for inter-annotator agreement

Empirical Analysis

Intentions	Examples and consequences	Def.	Mal.	Dox.
Endanger	May cause human trafficking, reputational risk, physical threats, sexualized misrepresentation, or hypocrisy	N	Y	Y
Scare, distress, or panic	Swatting or pushing activists offline by intimidating, alarming, or blackmailing	N	Y	Y
Defame or denigrate	Presenting disinformation, rumors, or real and private info about celebrities or public figures to discredit them	N	Y	Y | N
Digital vigilantism	To get the targets fired from their jobs, shamed in front of their neighbors, run out of town, or encourage reform or remorse of hate groups, discrimination, racism, sexism, and homophobia	Y	Y	Y | N
Doxing report	Describes, promotes/accuses someone about a doxing episode that happened (by quoting, replying, or indicating)	Y | N	Y | N	Y
Help seeking	Disclosing sensitive info about relatives by getting or pretending to be scared or worried about them	Y | N	Y | N	Y
Self-Protection	Public denunciations against misogynist trolls	Y	Y | N	Y
Self-Disclosure	Revealing the author as the bully, victim, defender, bystander, assistant, or re-enforcer	Y | N	N	N

Common motivations and intentions behind public sensitive information disclosures. “Y,” “N,” “Def.,” “Mal.,” and “Dox.” stand for “Yes,” “No,” “Defensive,” “Malicious,” and “Doxing” respectively. We use “Y | N” for different situations in which a category may or may not have a specific characteristic.

Empirical Analysis

Attributes	Malicious | Doxing	Benign | Self-Disclosure	Mal./Ben.
Total samples	2,135	996	2.14
Unique users	1,681	507	3.32
Mean status (tweet) count	20,928	85,625	0.24
Network characteristics			
No followers	1.19 %	2.56 %	0.47
No friends (followings)	0.60 %	1.97 %	0.31
Profile characteristics			
No Favorites	0.71 %	3.55 %	0.20
No location specified	26.47 %	28.80 %	0.92
No profile banner	2.86 %	15.58 %	0.18
No URL in bio	61.87 %	61.14 %	1.01
Customized profile theme	17.67 %	33.14 %	0.53
Use the default profile image	2.26 %	2.96 %	0.76
$\|\|Name\|\| < 3$	4.40 %	2.56 %	1.72
$\|\|Name\|\| > 20$	17.01 %	14.40 %	1.18
Have less than 10 tweets	1.07 %	3.75 %	0.29
Have less than 100 tweets	5.06 %	10.45 %	0.48
Created since 2019	92.81 %	52.66 %	1.76
Have verified badge	1	2	0.50

Analysis of outliers and individual user attributes. "$||Name|| < 3$" stands for the number of unique accounts that have less than 3 characters in their names. The number of unique users having a specific attribute is normalized by the total number of unique users in the corresponding class.

Automated Detection

Automated detection (binary classification) of :

Positive Class (+) : Doxing and malicious disclosure

Negative Class (-) : Self-disclosure and non-sensitive

Automated Detection

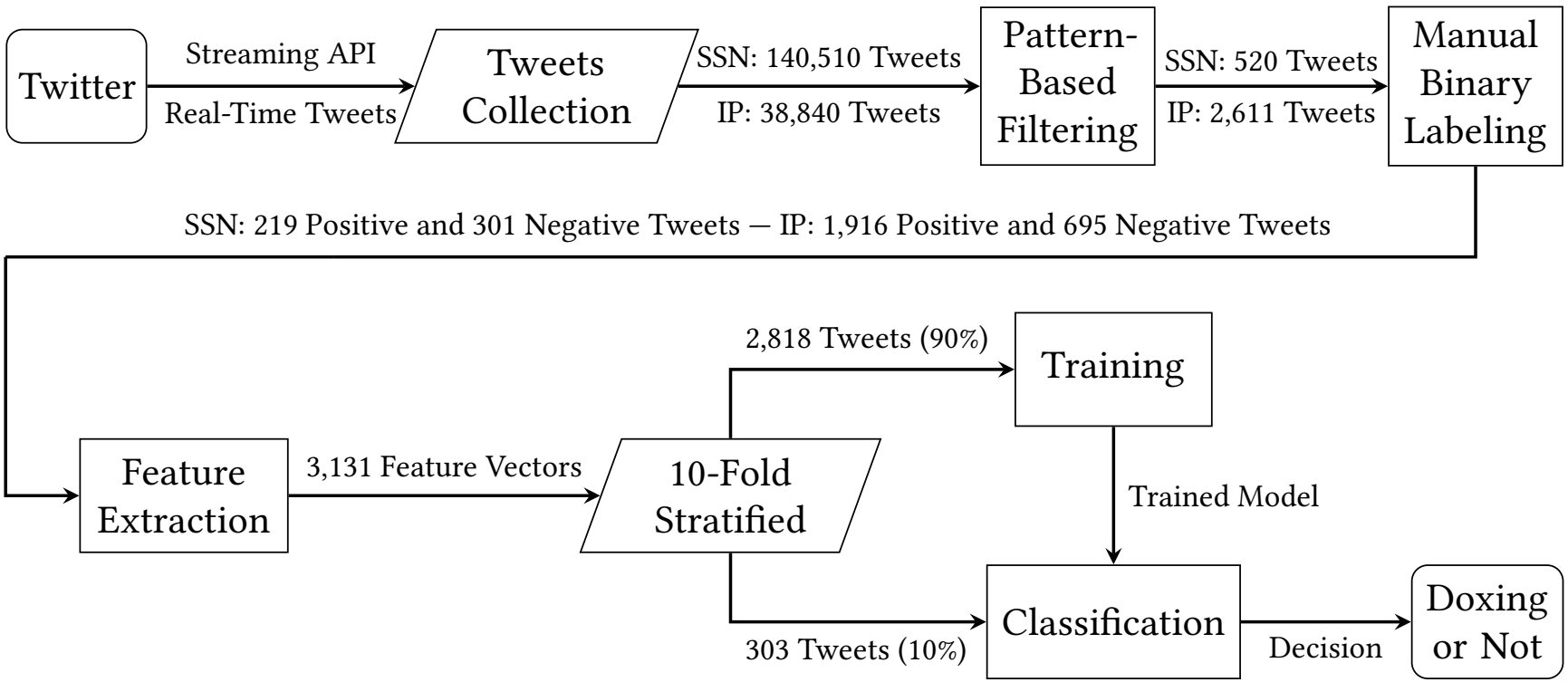

Fig. 1. Overview of our dataset creation and enrichment pipeline, combined with our experimental setup. Positive tweets are the ones that are doxing or malicious disclosures and negative tweets are the non-doxing tweets and self-disclosures.

Automated Detection

- Map textual data into numbers (understandable for computers)
- Preprocessing : Removing noise from tweets
  - Written in hurry
  - Spellings errors
  - Many redundant characters, emojis, emoticons
- Initial filtering
  - SSN : [1-899]–[1-99]–[1-9999]
    Not 666 – * * – * * * *
  - IP Address : [0-255].[0-255].[0-255].[0-255]
    Not 127.0.0.*, 192.168. * .*, 0.0.0.0, 8.8.8.8

Automated Detection

Feature Extraction Methods

- Heuristics : String-matching, one-hot encoding
  - "[Fail2Ban] POSTFIX-neelix"
  - Username + longitude+ lattitude + IP
- Classical word embedding : GloVe
- Contextualized string embedding :
  - Character-level embedding which captures word semantics in the context and produces different embeddings for polysemous words (to get : to understand/to arrive/to procure) based on their usage
  - Flair contextualized string embeddings
  - 2048 features per token (word)
- Document pool embedding : Generate a single representation for the whole tweet

PennState

Automated Detection

Heuristics

Type of heuristic	Strings
Common phrases	*left me no choice, this you, deactivate luv, dox, cyberbully, loser, watch out, scare, dumb, your ip address is, your ssn is, blacks, black people, nigga, hate, fuck, bitch, shut up, delete this or, delete the video or, warned you, delete the image or, death, troll, ass, shit, delete your video, i have your ip address, your photos will be posted*
Invalid-looking SSNs	*111-11-1111, 222-22-2222, 333-33-3333, 444-44-4444, 555-55-5555, 666-66-6666, 777-77-7777, 888-88-8888, 999-99-9999, 123-45-6789, 696-96-9696, 420-69-6969, 420-69-6669, 420-69-6666, 420-69-1488 420-69-1337, 420-69-8008, 420-69-1312, 420-69-1313, 420-69-1234, 420-69-2001, 420-69-1969, 420-69-1738, 078-05-1120*

Phrases and invalid-looking SSNs we used as our heuristics

Automated Detection

Possible Figurative Meanings

Indicators	Possible figurative meaning
xxx-xx-xxxx	All the digits are the same
123-45-6789	The sequence of all digits
078-05-1120	The SSN used and widely distributed by the Woolworth wallet manufacturer. [1]
69	A sex position
420	Slang for Marijuana and smoking pot
666	Known as the number of the beast or Devil's number
1234	The sequence of numbers (if used along with 420 and 69)
1312	Its digits represent the first three alphabet letters, "ACAB," which is an acronym used as a political slogan by those who are opposed to the police and stands for "All Cops Are Bastard"
1313	Consists of a pair of 13s which is sometimes referred to as the number of bad (or good) luck or new beginnings
1337	Represents the term "LEET"
1488	The 14-word slogan of white supremacists and 8 stands for 'H', the 8^{th} letter in alphabet, and 88 is a code representing the initials for "Heil Hitler"
1738	Remy Martin Cognac
1969	A birth year in the 20^{th} century that ends with 69
2001	The year of the September 11 attacks
8008	A representation for the term "BOOB"

Possible figurative meaning of specific digits that may be indications of invalid SSNs and false disclosures.

1. https ://www.ssa.gov/history/ssn/misused.html

Automated Detection

Method	Features	Feat.	Pos.	Neg.	Train	Test
Heuristics	Table 5 strings and IP address heuristics	N/A	2135	996	N/A	3131
1-HotEH	One-hot encoded strings from Table 5	67	2135	996	2818	313
1-HotEH_Heuristics	One-hot encoded strings, overruling heuristics	67	2135	996	2818	313
Mean_GloVe_Twitter	Average of GloVe Twitter model word embeddings	200	2135	996	2818	313
DP_GloVe_Wiki	Document pool embedding of GloVe Wikipedia model	100	2135	996	2818	313
DP_FlairFW	Document pool embedding of Flair news forward model	2048	2135	996	2818	313
DP_FlairFW_Cleaned	Document pool embedding of Flair news forward model	2048	2132	897	2726	303
DP_FlairFW_Heuristics	Document pool embedding of Flair news forward model, overruling heuristics	2048	2135	996	2818	313
DP_FlairFW_GloVe_Wiki	Document pool embedding of Flair and GloVe models	2148	2135	996	2818	313

Configurations used for different detection approaches. “Feat.,” “Pos.,” and “Neg.” stand for “Features,” “Positive samples,” and “Negative samples” respectively. Note that the train and test sizes are the average sizes per fold in a 10-fold stratified cross validation, except for the first row which does not require any training. Also, the cleaned dataset has 94 tweets less than others in totals which are the invalid-looking SSNs that are removed. Furthermore, the number of features only represents the features used for an automated classification task and does not include the strings used as our heuristics.

PennState

Automated Detection

Method	TPR	TNR	FPR	FNR	Acc. %	Prec. %	Rec. %	F1 %
Heuristics	0.20	0.62	0.38	0.80	33.47	53.22	20.14	29.22
1-HotEH	0.99	0.11	0.90	0.01	71.10	70.42	99.34	82.42
1-HotEH_Heuristics	1.00	0.10	0.90	0.00	71.19	70.35	99.81	82.53
Mean_GloVe_Twitter	0.97	0.92	0.08	0.03	95.37	96.19	97.05	96.62
DP_GloVe_Wiki	0.97	0.93	0.07	0.03	95.46	96.76	96.58	96.67
DP_FlairFW	0.94	0.87	0.13	0.06	91.25	92.58	93.64	93.11
DP_FlairFW_Cleaned	0.97	0.96	0.04	0.03	96.86	98.16	97.37	97.76
DP_FlairFW_Heuristics	0.98	0.90	0.10	0.03	95.05	95.37	97.47	96.41
DP_FlairFW_GloVe_Wiki	0.97	0.95	0.05	0.03	96.61	97.74	97.28	97.51

Comparison of different detection approaches. “Acc.,” “Prec.,” “Rec.,” and “F1” stand for “Accuracy,” “Precision,” “Recall,” and “F1-score” respectively.

Discussion and Limitations

1. All the sensitive information that we find, manually or automatically, are purported
2. We cannot verify whether the information is real and belongs to the claimed identity.
3. We cannot differentiate between doxing and other malicious disclosures.
4. Suspicious tweets and accounts that have been removed or suspended later
5. Any normal Twitter user have access to all that sensitive private information before Twitter even gets to take any actions against them.

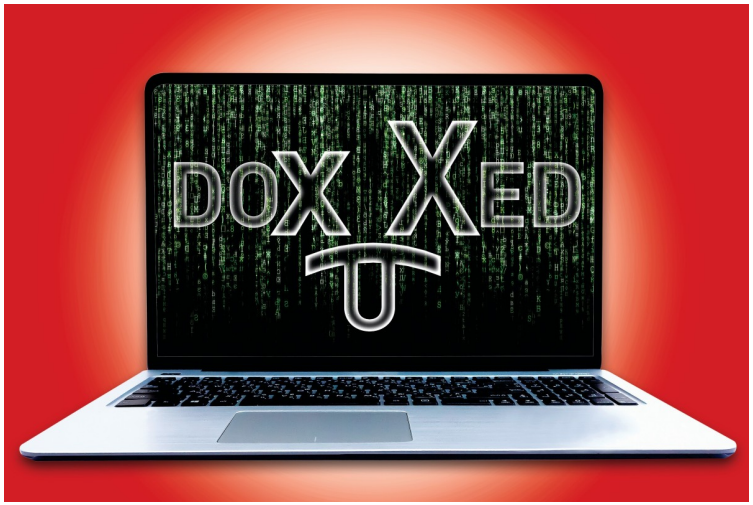

Image : scmp.com/magazines/post-magazine/short-reads/article/3036663/doxxing-powerful-weapon-hong-kong-protests-had

Next Steps

1. Perform the analyses on a larger scale with more diverse tweets
2. Can we observe other cyberbullying habits among doxers ?
3. Serial doxers : Do doxers disclose private information about different victims ?
4. What types of doxed information attract more attention and get viral faster/more ?
   - About a misbehavior ?
   - About a celebrity or a public figure ?
5. Can we characterize and rate users based on their individual and inter-personal attributes to increase the confidence of automated doxer/doxing detection ?
6. Analysis of cyberbullies
   - Whether they get suspended
   - How long does it take from Twitter to identify/restrict abusive contents/users
   - What impacts does Cyberbullying have on cyberbullies over a time period. E.g., temporal sentiment analysis (unethical behavior consciously or unconsciously contributes to destroying one's own positive self-image)
   - What impacts do cyberbullies have on the targets/victims (conversation analysis)

PennState

References

Canfora, G., Di Sorbo, A., Emanuele, E., Forootani, S., and Visaggio, C. A.
A nlp-based solution to prevent from privacy leaks in social network posts.
In *Proceedings of the 13th International Conference on Availability, Reliability and Security* (2018), pp. 1–6.

Deodhar, L., Divakaran, D. M., and Gurusamy, M.
Analysis of privacy leak on twitter.
In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (2017), IEEE, pp. 1–6.

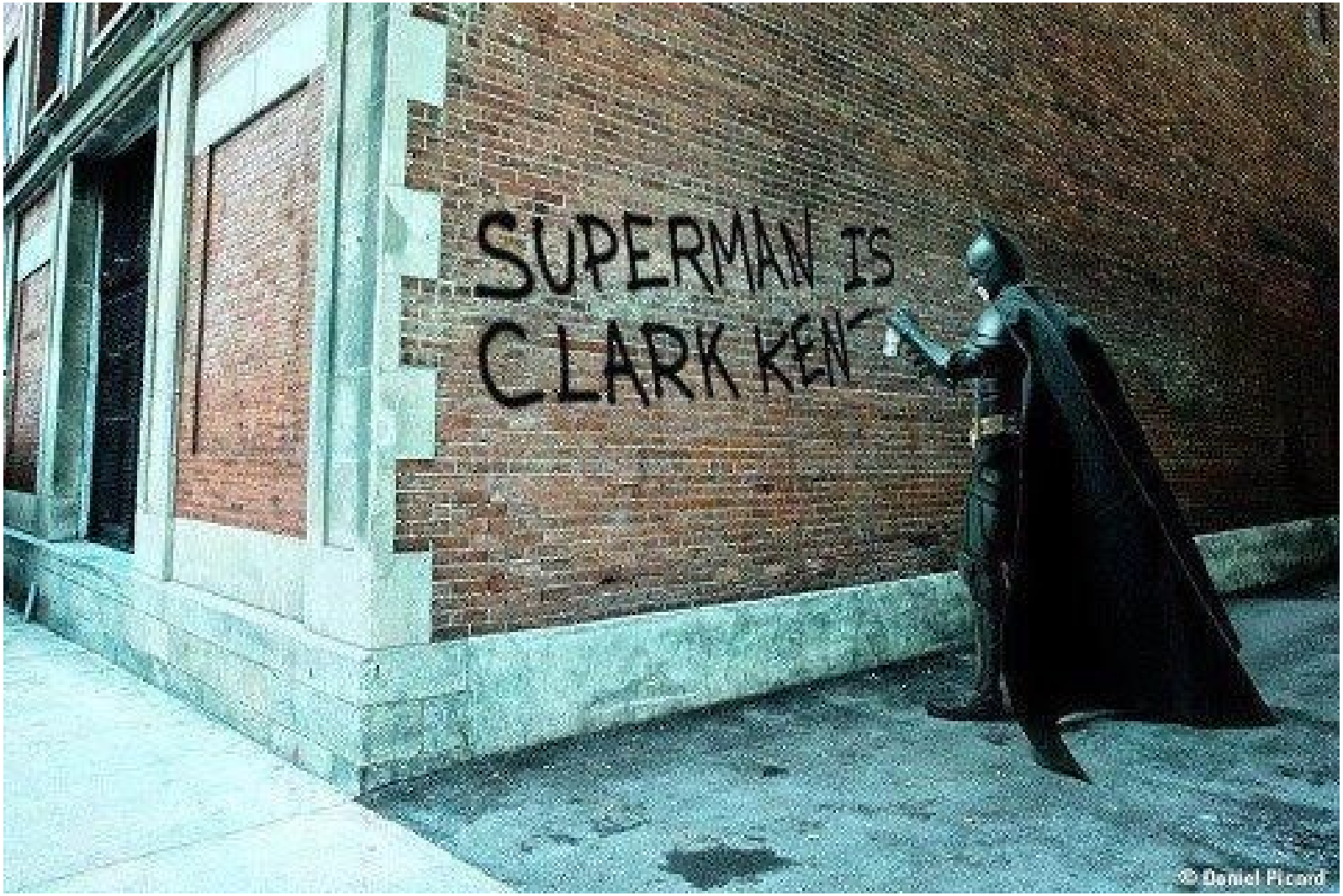

Image : cheezburger.com/9168162304/superhero-doxxing

PennState

Labeling Criteria

Doxing (TRUE)	**Non-doxing (FALSE)**
Sensitive disclosure about 2^{nd}/3^{rd} parties (info is connected to victim's identity by quoting them, mentioning their full names, usernames, etc. regardless of their potential intentions.	Self-disclosure
Promoting doxed info (report/reply/quote of a tweet that contains doxing ; even by the victim)	Does not target any specific/unique identity/person
Can be used to uniquely identify or physically locate a person	
There is no direction mention of other identities, but there are indicators (e.g., your SSN is . . .)	

Specific rules and criteria we provided to our annotators for labeling a sample of 100 tweets from the dataset to calculate the inter-annotator reliability and agreement.

PennState