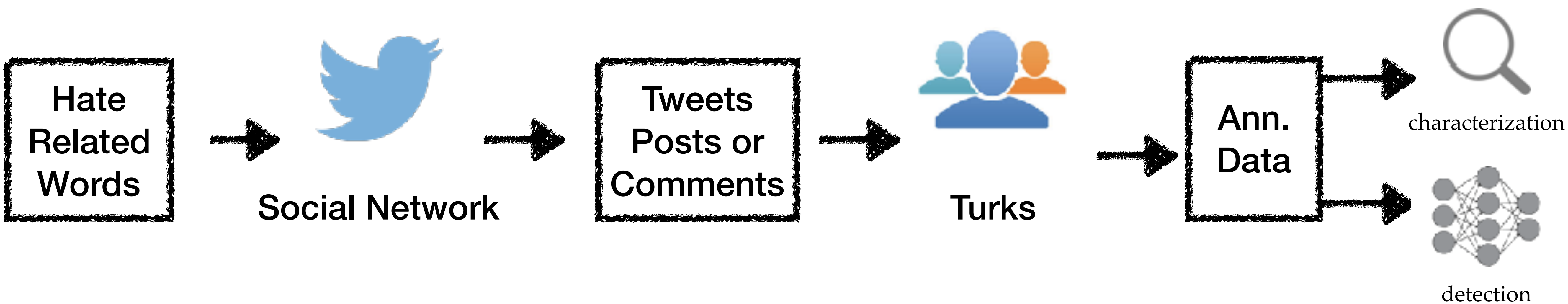
Like Sheep Among Wolves:
Characterizing Hateful Users on Twitter

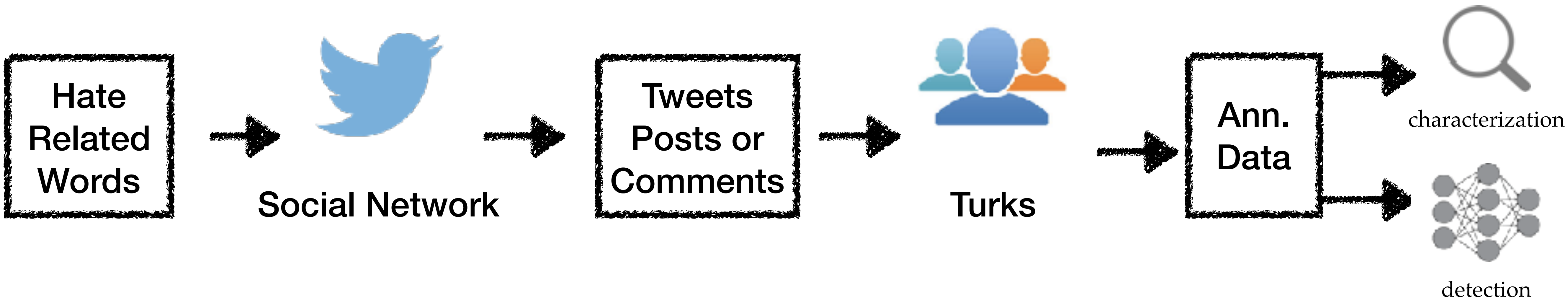
Manoel Horta Ribeiro
Pedro H. Calais
Yuri A. Santos
Virgílio A. F. Almeida
Wagner Meira Jr.

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- In recent years plenty of work was done on *characterizing* and *detecting* hate speech.

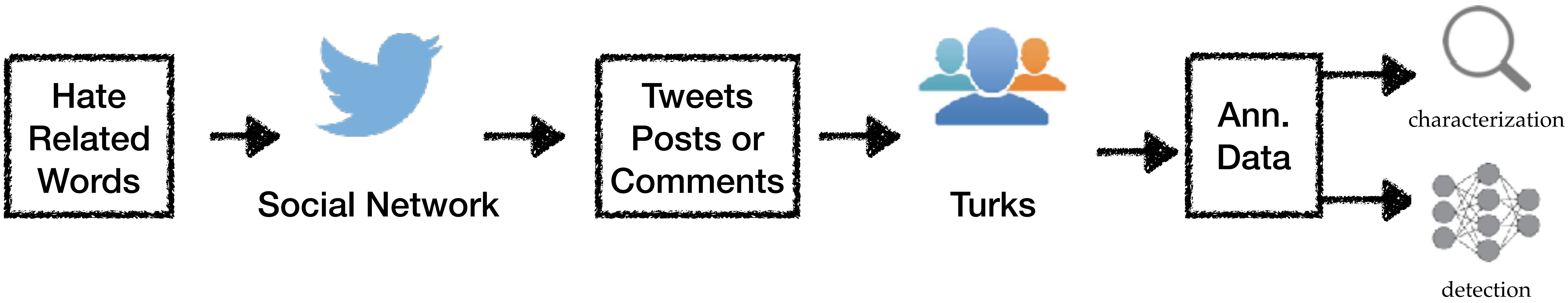


[Burnap and Williams 2017]
[Waseem and Hovy 2016]
[Davidson et al. 2016]



- the meaning of such content is often not self-contained;

> Time's up, you all getting what should have happened long ago



- hate speech != offensive speech

You stupid {insert racial slur here}

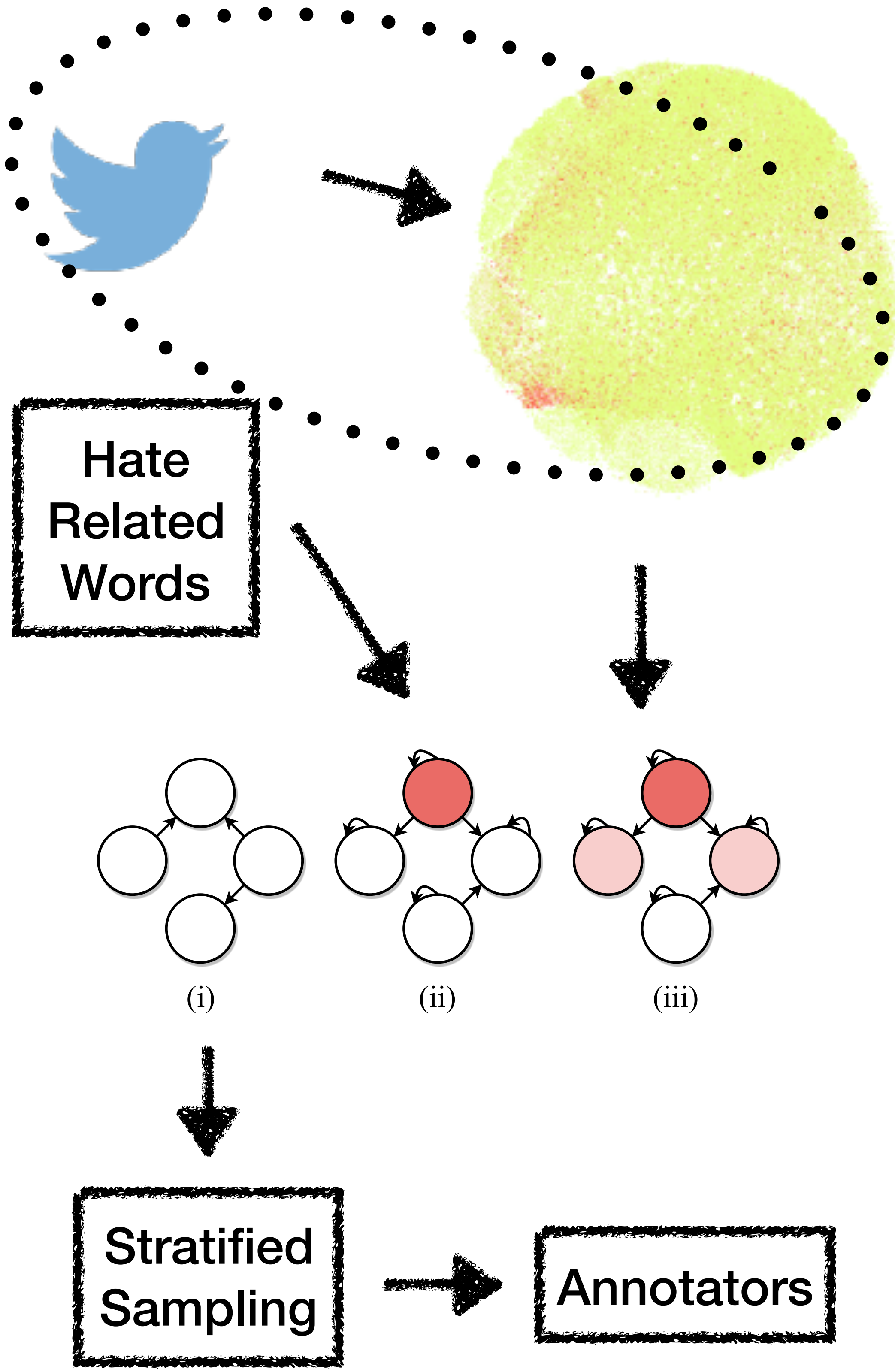
[Davidson et al. 2016]

- The previous work focuses on **content**, and has shortcomings related to **context**.

- Idea: change the focus from the *content*, to the *user*.

- Allows for more sophisticated data collection
- Give annotators context - not isolated tweets
- Richer feature space: activity, net. analysis

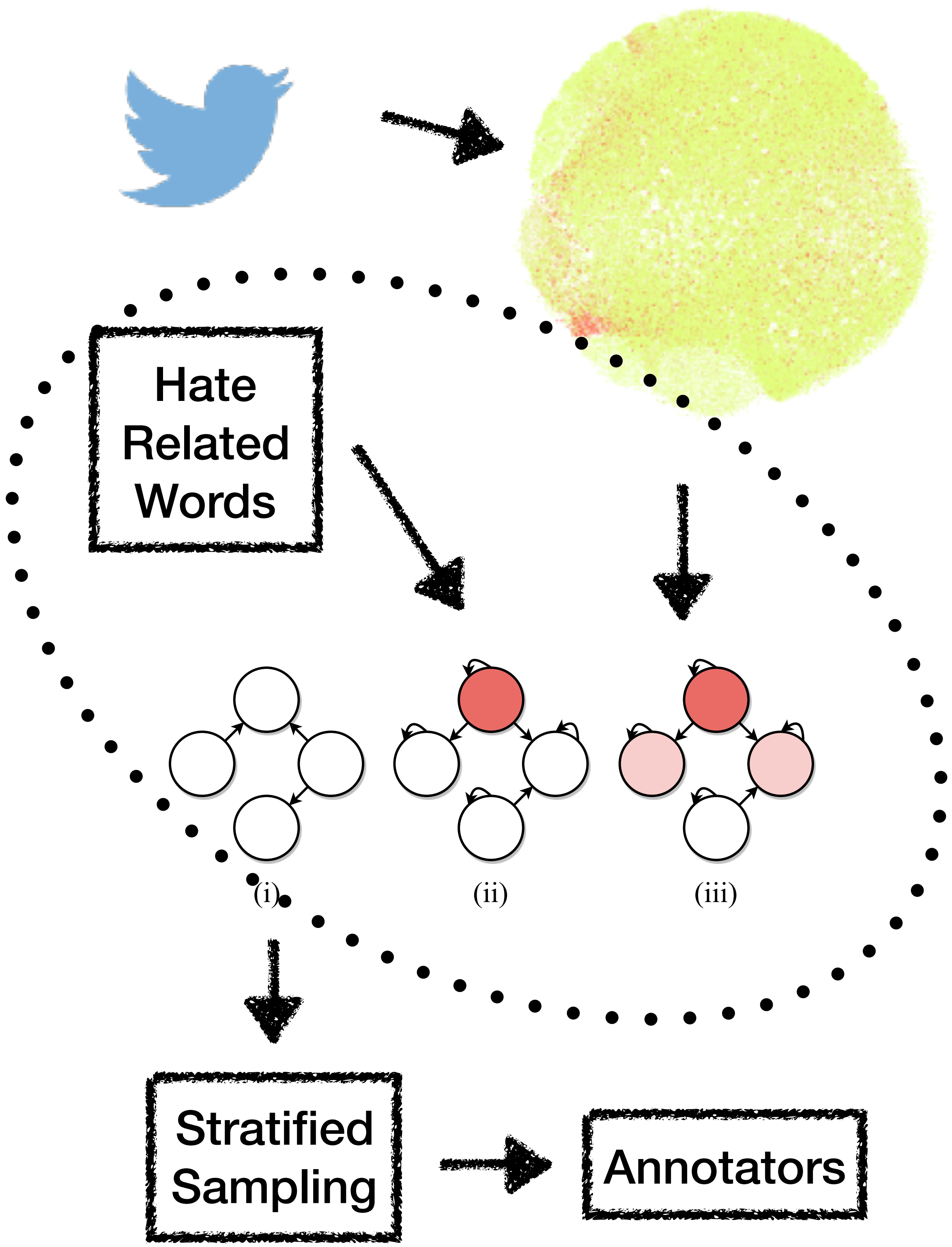
Data Collection



- We begin by sampling Twitter's retweet network. We employ a Direct Unbiased Random Walk (*DURW*) algorithm.
- Obtained 100,386 users, along with up to 200 tweets of their timelines.

[Ribeiro, Wang and Tosley 2010]

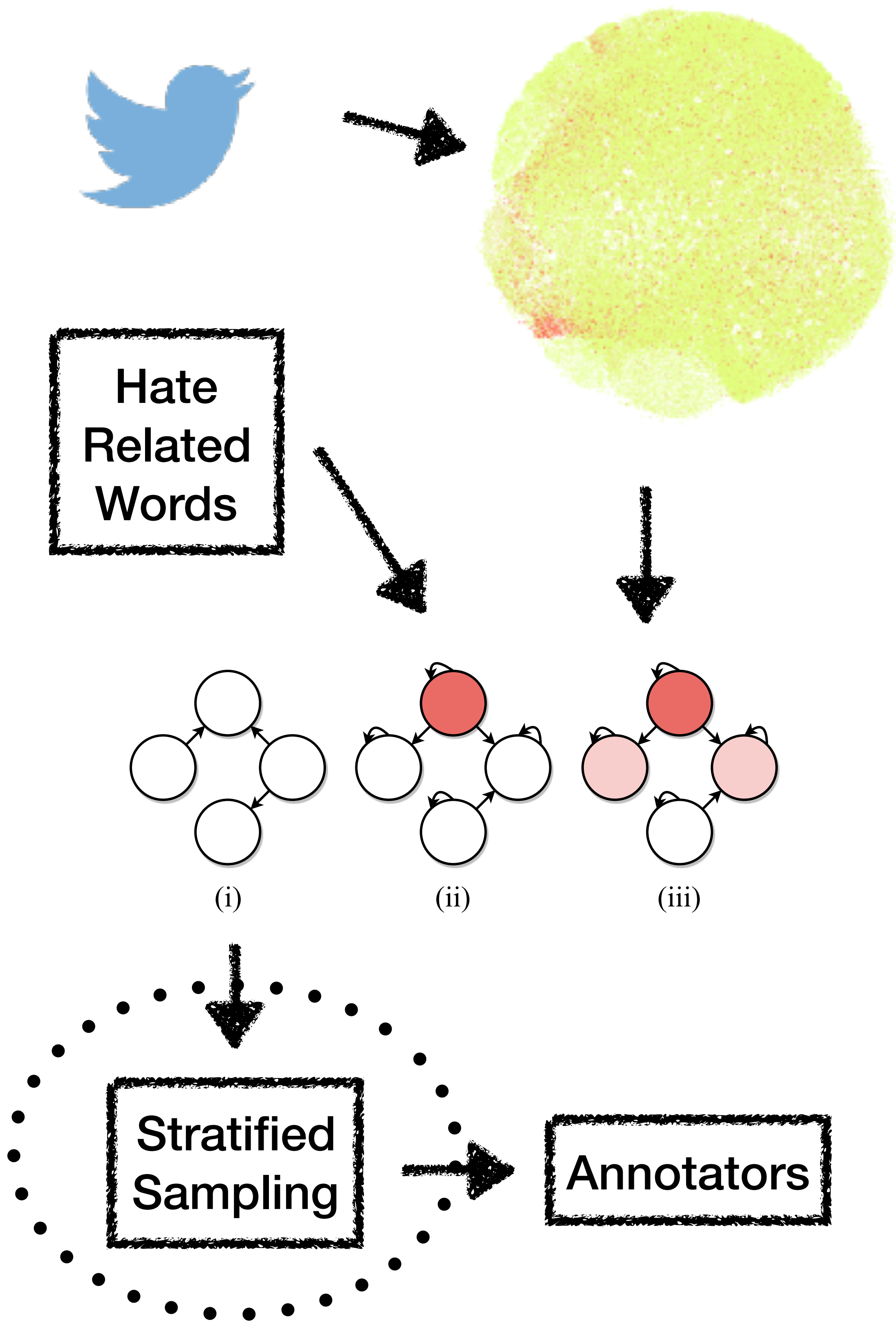
Data Collection



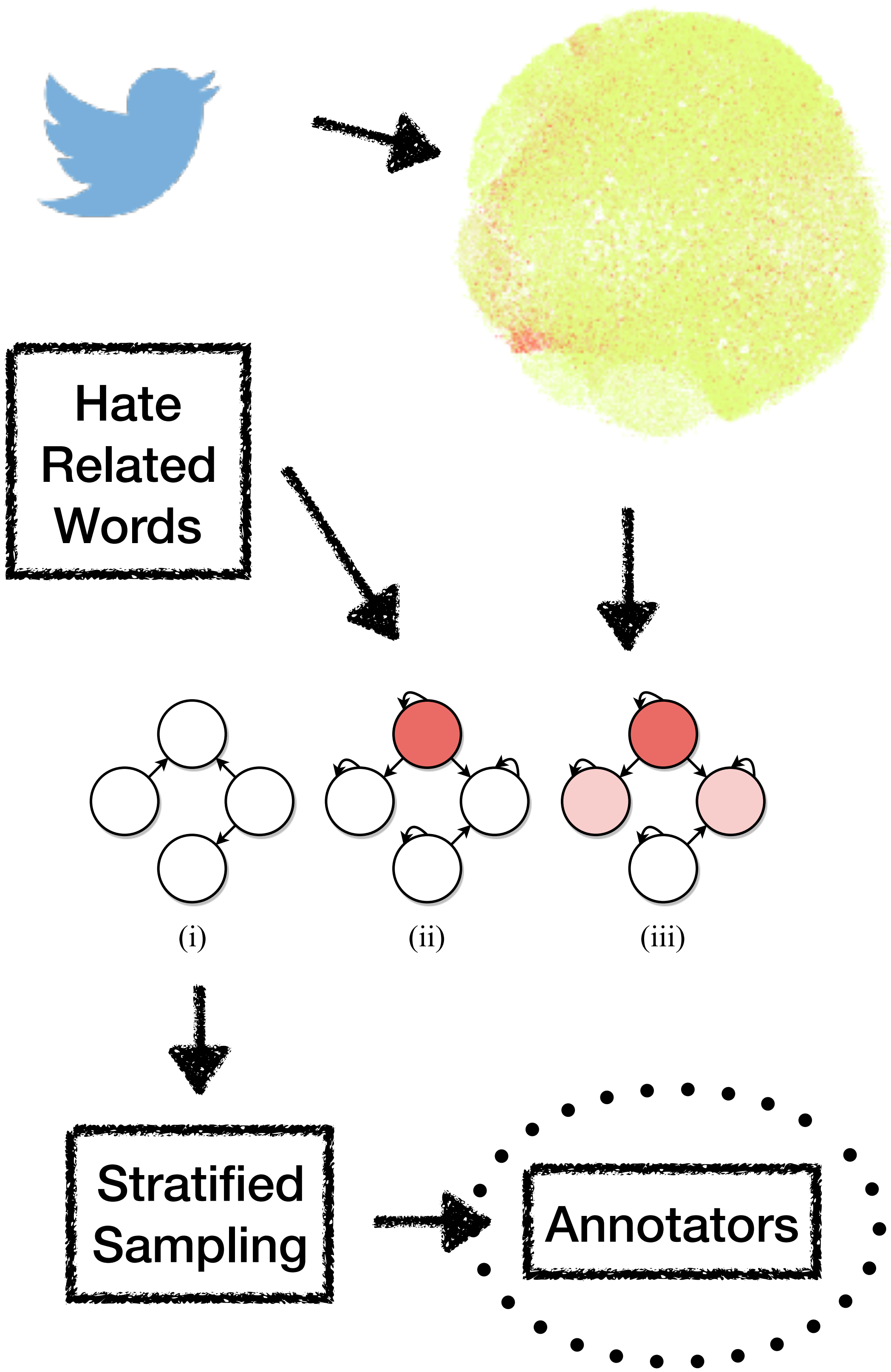
- Given the graph, we employ a hate related lexicon, tagging the users that employed the words.
- We use this users as seeds in a diffusion process based on DeGroot's learning.

[Golub and Jackson 2010]

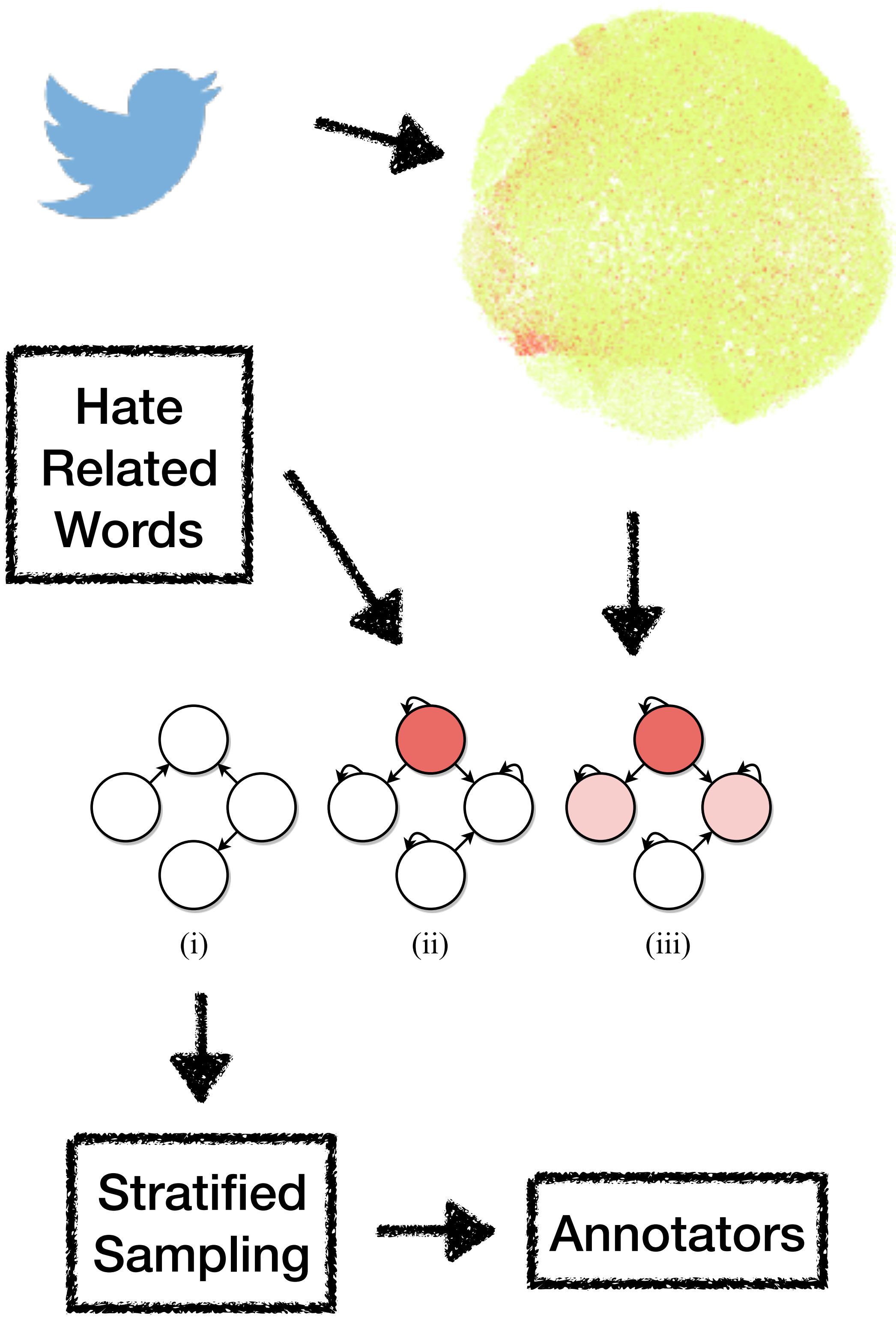
Data Collection



- After that, we have a real number in the range [0,1] associated with each individual in the graph.
- We then perform stratified sampling, obtaining up to 1500 users in the intervals [0,.25), [.25,.5), [.5,.75), [.75,1).



- We ask annotators to determine if users are hateful or not. They were asked to use Twitter's hateful conduct guideline.
- 3-5 annotators / user, obtained 4972 annotated users. 544 were considered hateful



- Lastly we also collect the users who have been suspended 4 months after the data collection.

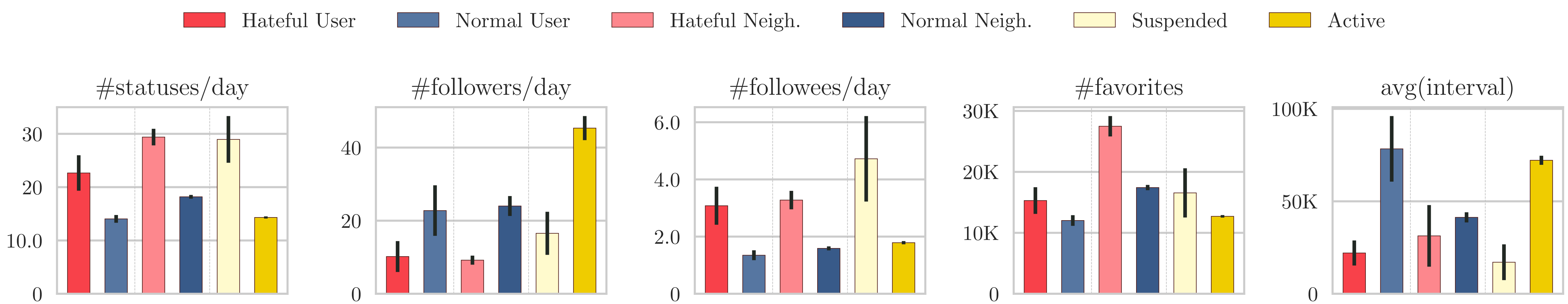
- We use Twitter’s API and obtain 686 suspended users.

Results

Hateful User | Normal User | Hateful Neigh. | Normal Neigh. | Suspended | Active

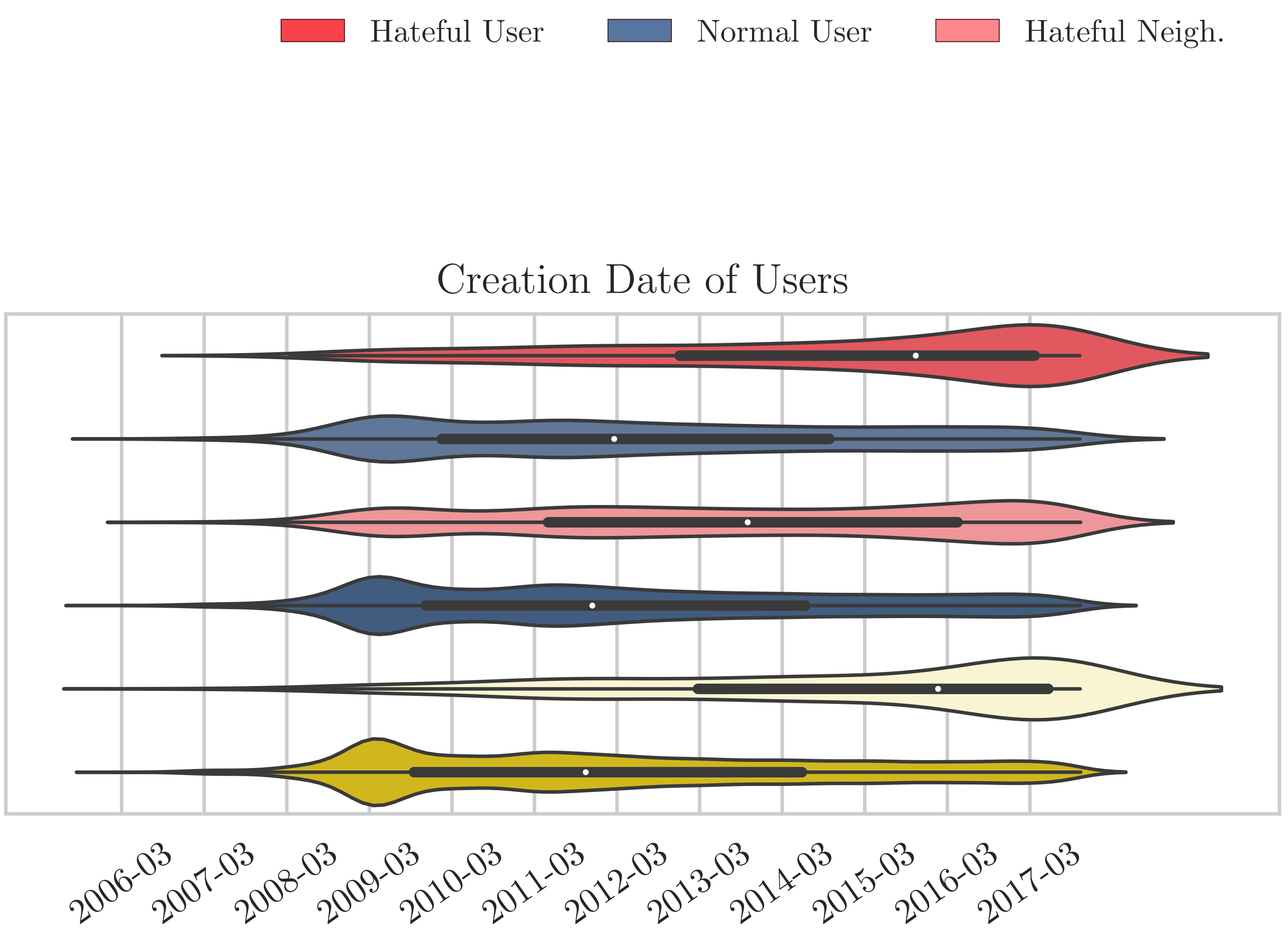
- We analyze how hateful and normal users differ w.r.t. their activity, vocabulary and network centrality.
- We also compare the neighbors of hateful and of normal users, and suspended/active users to reinforce our findings.
- We compare those in pairs as the sampling mechanism for each of the populations is different.
- We argue that each one of these pairs contains a proxy for hateful speech in Twitter.

Hateful Users are power users



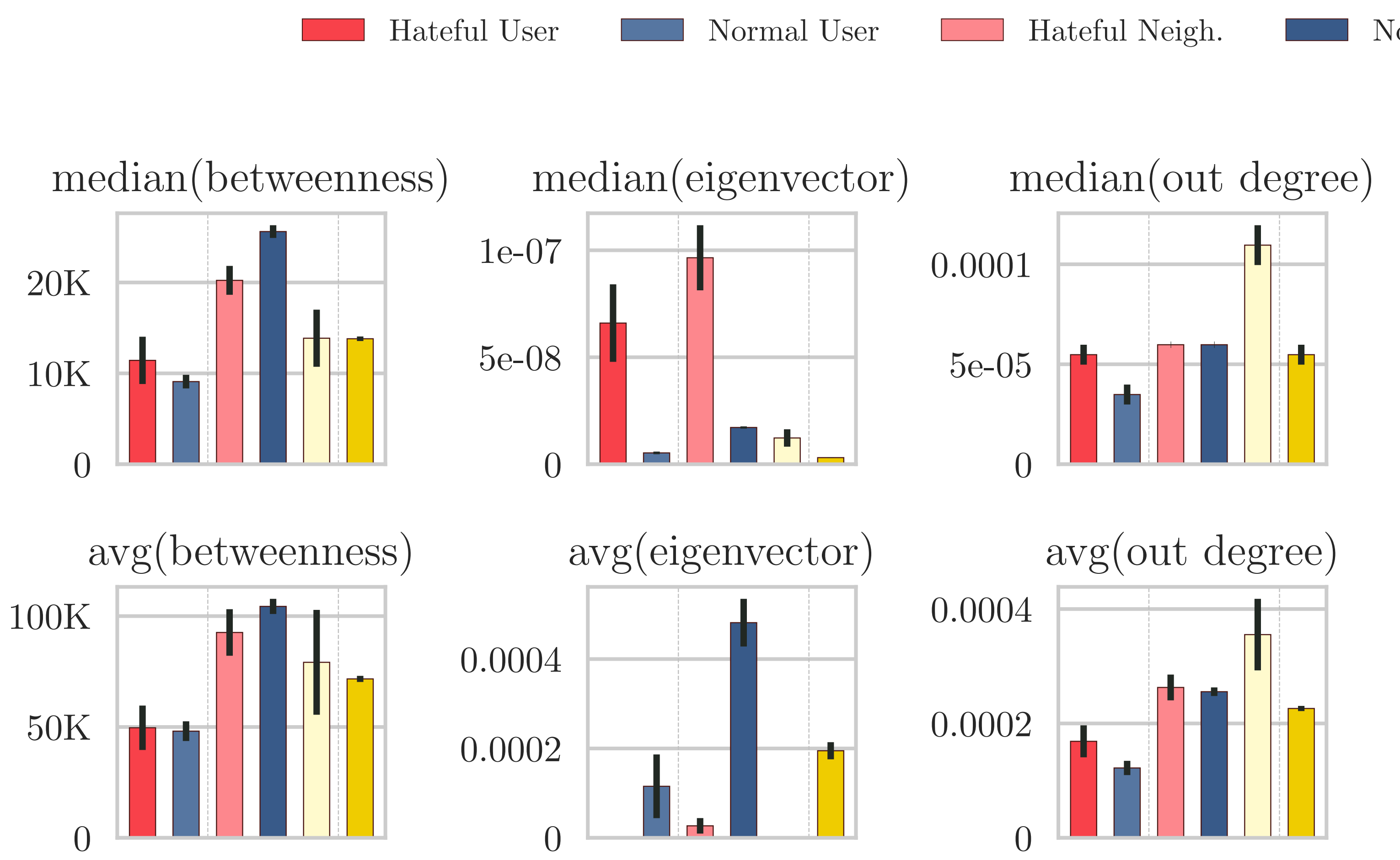
- Hateful users tweet more, in shorter intervals, favorite more tweets by other people and follow others more (p-values <0.01).
- We observe similar results when comparing their neighborhood and when comparing active vs. suspended users.

Hateful users have newer accounts



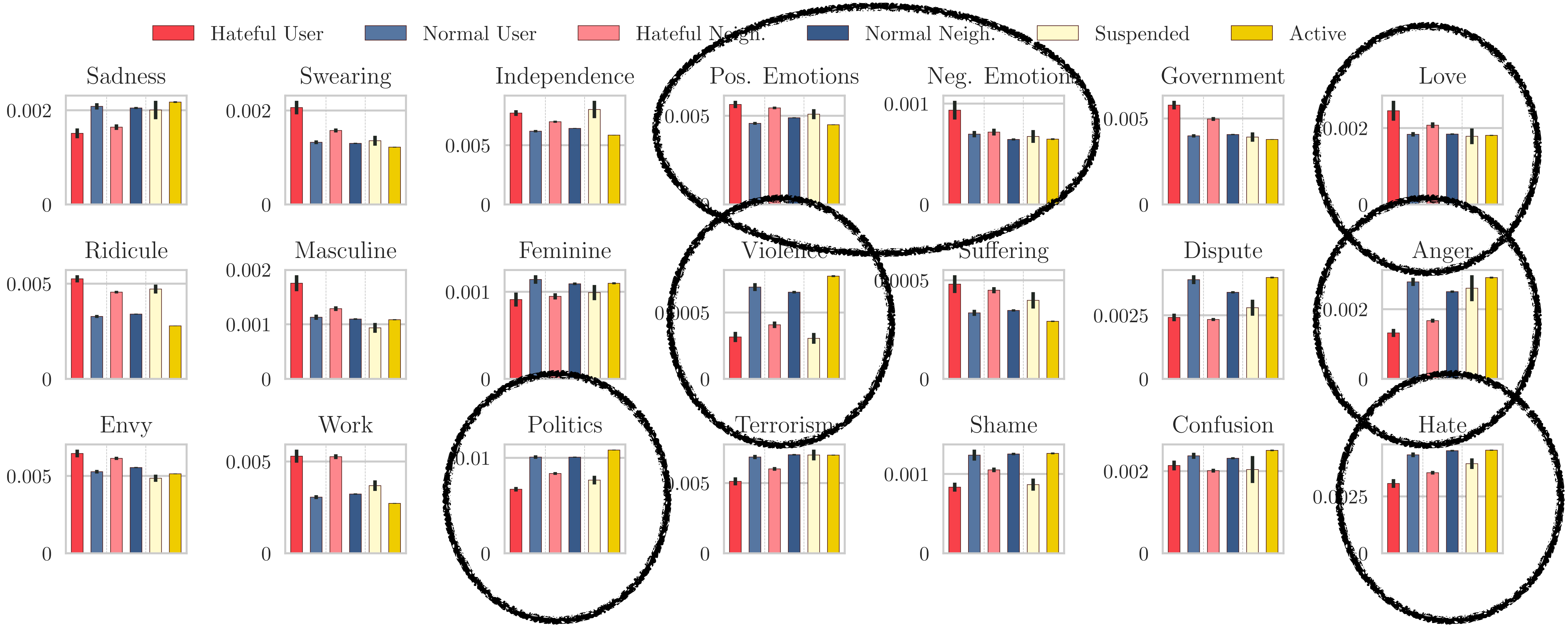
- Hateful users were created later than normal ones (p-value < 0.001).
- A hypothesis for this difference is that hateful users are banned more often due to Twitter's guidelines infringement.

The median hateful user is more central



- Median hateful user is more central in all three measures.
- Average hateful user isn't, deformed by very influential users.

Hateful users use non-trivial vocabulary



- Average values for the usage of *EMPATH* lexical categories.

Future Stuff / Discussion

Node Type	(%)	Node Type	(%)
🔴 → 🔴	41.50	🔴 → 🔵	13.10
🔵 → 🔵	15.90	🔵 → 🔴	2.86
⚪ → ⚪	7.50	⚪ → 🟡	92.5
🟡 → 🟡	99.35	🟡 → ⚪	0.65

⚪ Suspended 🟡 Active

🔴 Hateful User 🔵 Normal User

- hateful users are 71x more likely to retweet another hateful user.
- suspended users are 11x more likely to retweet another suspended user.

Future Stuff / Discussion

		Hateful/Normal			Suspended/Active		
Model	Features	Accuracy	Recall	AUC	Accuracy	Recall	AUC
GradBoost	user+glove	84.6 ± 1.0	76.7 ± 2.4	88.4 ± 1.3	81.5 ± 0.6	78.9 ± 1.7	88.6 ± 0.1
	glove	84.4 ± 0.5	77.2 ± 2.1	88.4 ± 1.3	78.9 ± 0.7	77.7 ± 1.6	87.0 ± 0.5
AdaBoost	user+glove	69.1 ± 2.4	84.6 ± 1.9	85.5 ± 1.4	70.1 ± 0.1	84.4 ± 3.6	84.3 ± 0.5
	glove	69.1 ± 2.5	84.8 ± 1.8	85.5 ± 1.4	69.7 ± 1.0	83.0 ± 0.3	82.7 ± 0.1
GraphSage	user+glove	$\mathbf{90.9 \pm 1.1}$	84.6 ± 6.0	$\mathbf{95.4 \pm 0.2}$	$\mathbf{84.8 \pm 0.3}$	$\mathbf{85.6 \pm 5.4}$	$\mathbf{93.3 \pm 1.4}$
	glove	90.3 ± 1.9	$\mathbf{85.1 \pm 7.6}$	94.9 ± 2.6	84.5 ± 1.0	85.5 ± 3.9	93.3 ± 1.5

- We can also bring the idea of bringing the focus to the user for the task of classification.
- Features:
  - *GloVe* vectors for the tweets (average);
  - Activity / Network centrality attributes;
- **Beyond new features, we may use the very structure of the network in the classification task.**

Summary

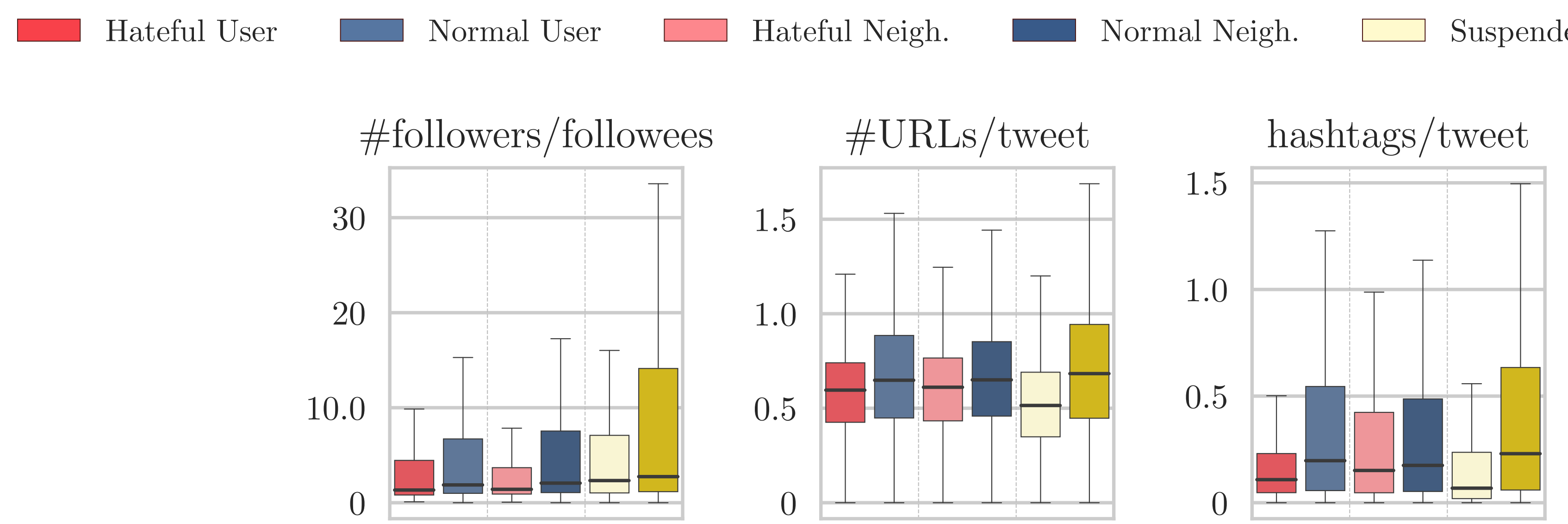
1. Proposed changing the focus from content to user;
2. Proposed a data collection method with less bias towards a specific lexicon;
3. Observed significant differences w.r.t. activity, lexicon and net centrality between hateful and normal users.
4. Showed how the network structure of users can be used to improve detecting hateful and suspended users.

github manoelhortaribeiro
twitter manoelribeiro
mail manoelribeiro at dcc.ufmg.br

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INCIPIT VITA NOVA
7 DE SETEMBRO DE 1927

Hateful users don't behave like spammers



- We analyze metrics that have been used to detect spammers.

- Hateful user in our dataset do not seem to be abusing hashtags or mentions, and do not have higher ratios of followers per followees.