Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter

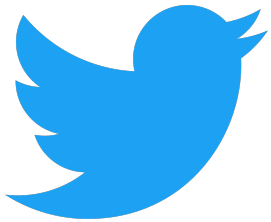
Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn
Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali

Hypertext
Prague, Czech Republic, 2017

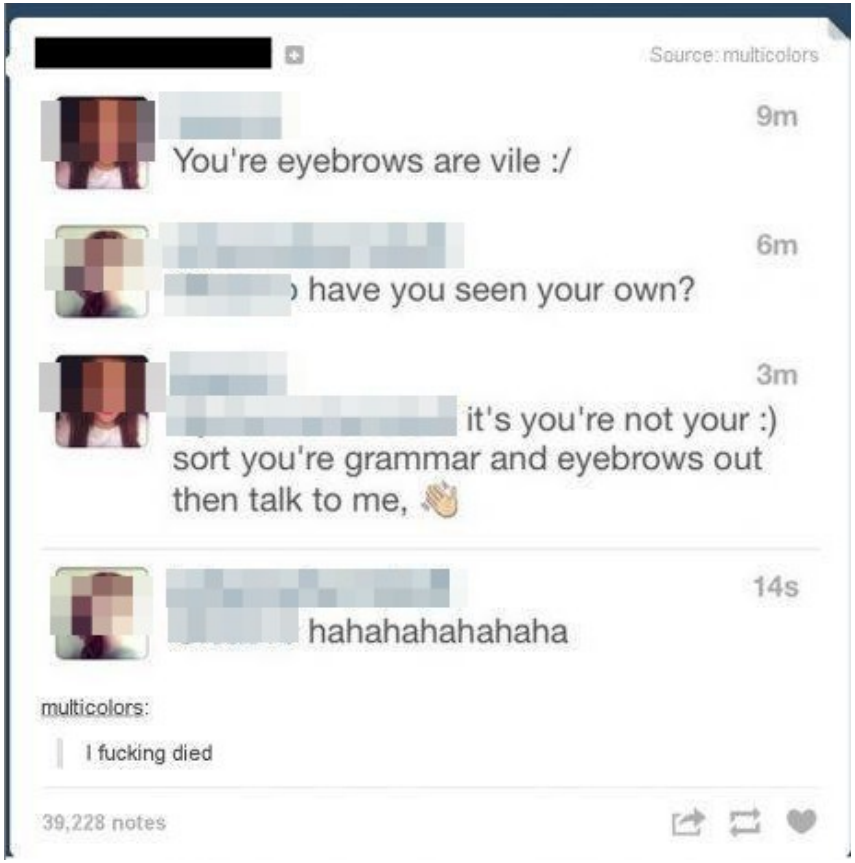
Social Networking Services

Social networking applications contain user profiles, variety of resources, and activities.

Twitter



- A microblogging service
- Sharing of up to 140-character messages
- Sharing of any kind of content



Cyberbullying vs. Cyberaggression

- **Cyberaggression:** purposefully saying or doing something to hurt someone once
- **Cyberbullying:** intentionally aggressive behavior, repeated over time, that involves an imbalance of power

Goal No1

What distinguish Abusers from Typical twitter users?

Open Questions

- What are the characteristics of abusers and typical twitter users?
- How users' emotional and activity characteristics can be used for distinguishing among different users' behaviors?

Goal No2

How the Suspension and Deletion mechanism works on Twitter?

Is this goodbye?

Are you sure you don't want to reconsider? Was it something we said? Tell us.

Before you deactivate [redacted], know this:

- **We will only retain your user data for 30 days** and then it will be permanently deleted. You can reactivate your account at any point within 30 days of deactivation by logging back in.
- You don't need to deactivate your account to change your username or Twitter URL. You can change it on the settings page. All @replies and followers will remain unchanged.
- If you want to use this account's username or email address on another Twitter account, chang[redacted] deactivate. Until the user data is permanently deleted, that information won't be available for us
- Your account should be removed from Twitter within a few minutes, but some content may be v[redacted] twitter.com for a few days after deactivation.
- We have no control over content indexed by search engines like Google.

Deactivate [redacted] Cancel

Open Questions

- What is the twitter account status and how do we measure it?
- What are the characteristics of suspended users and users who deleted their Twitter account?
- What are the characteristics of users who remain active on Twitter, but should have been suspended?
- Can we emulate the Twitter account suspension mechanism?

Datasets

Crawling from June to August 2016:

- **Baseline:** 1M random tweets
- **Hate-related:** 650K tweets based on 309 bully- and hate-related hashtags

309 hashtags: **#GamerGate** and 308 co-appeared ones

Gamergate Controversy

- A coordinated campaign of harassment in the online world
- It invovles sexism, feminism, and "social justice" and takes place on social media like Twitter

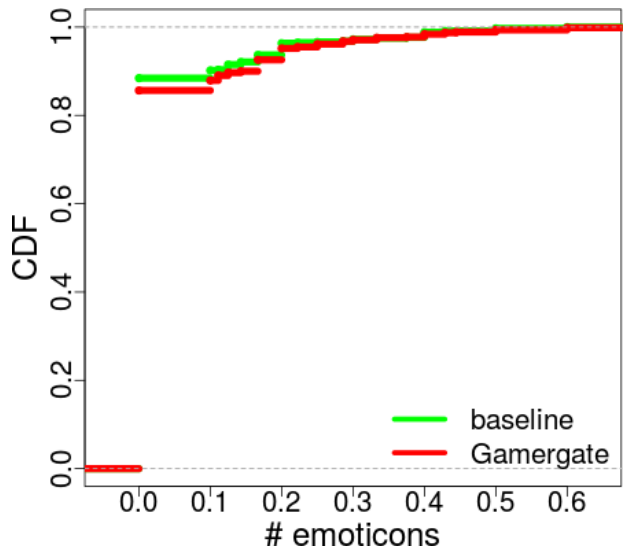
Gamergate controversy provides us a unique point of view into online harassment campaigns

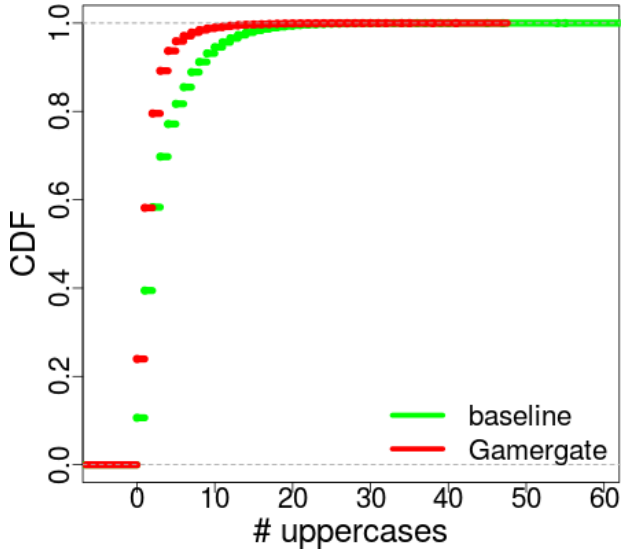
Goal 1

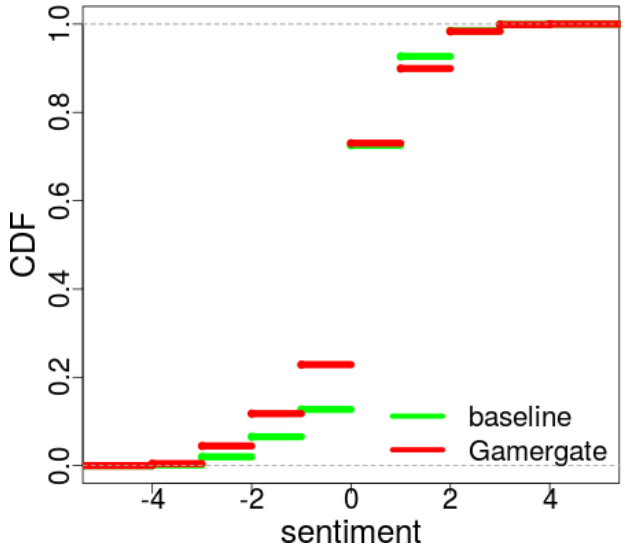
Considered Axes

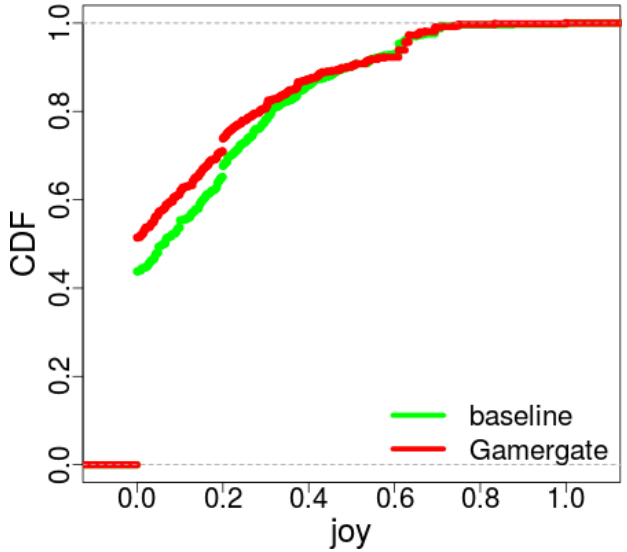
- **Emotional characteristics**: sentiment, emotions, offensive, emoticons, uppercase
- **Activity characteristics**: account age, # posts / lists / favorites, mentions, followers, friends

Emotional Characteristics









Emotional Characteristics - Findings

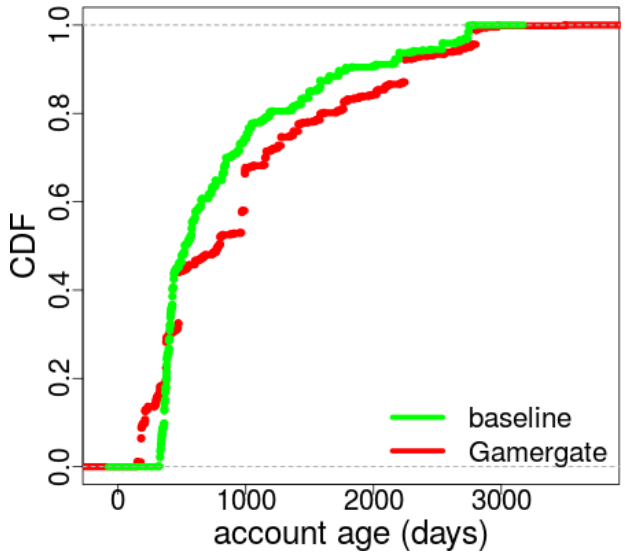
Emoticons and "shouting" by using all capital letters:

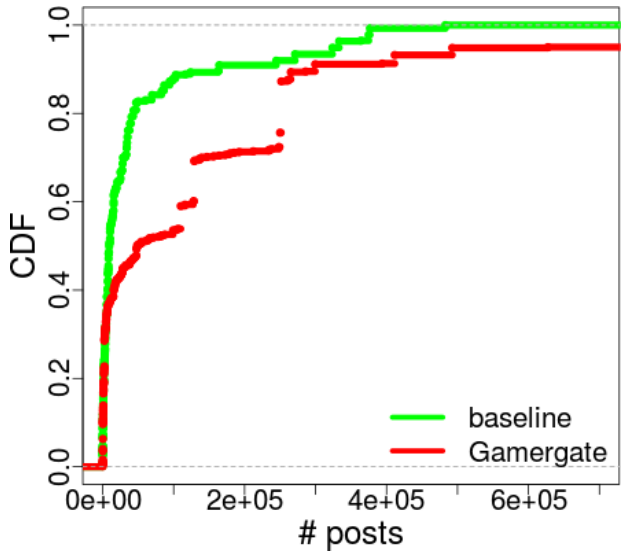
- GGers and baseline users use emoticons at about the same rate
- GGers tend to use all uppercase less often than baseline users

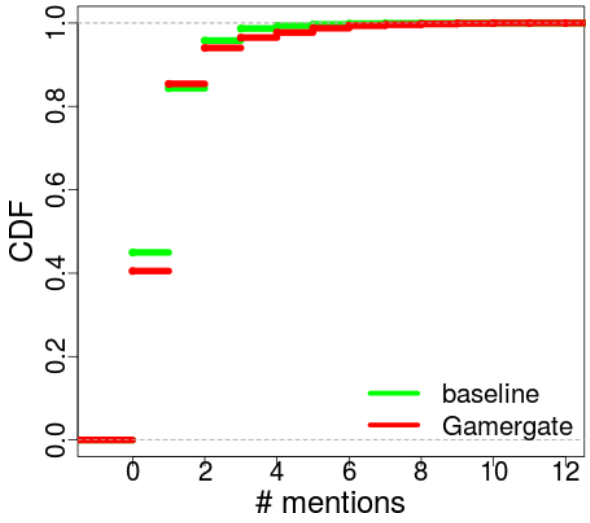
Sentiment, Emotion, and Offense

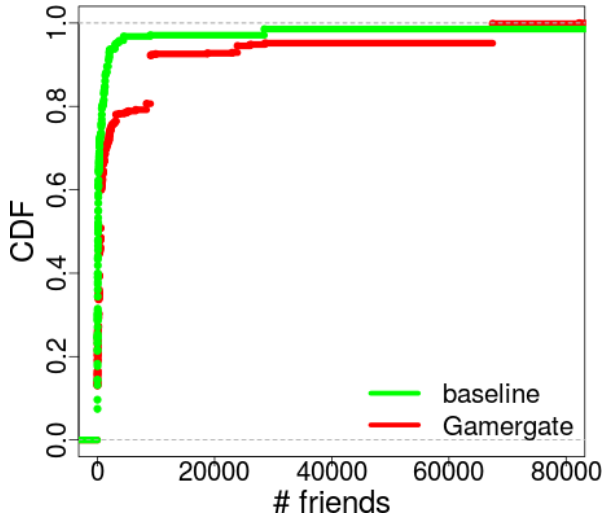
- GGers post tweets with a generally more negative sentiment
- GGers use more hate words than random users (Hatebase database)
- GGers and baseline users do not differ substantially in a variety of emotions: anger, disgust, fear, sadness, surprise
- GGers are less joyful $->$ they are not necessarily angry, but they are apparently not happy

Activity Characteristics







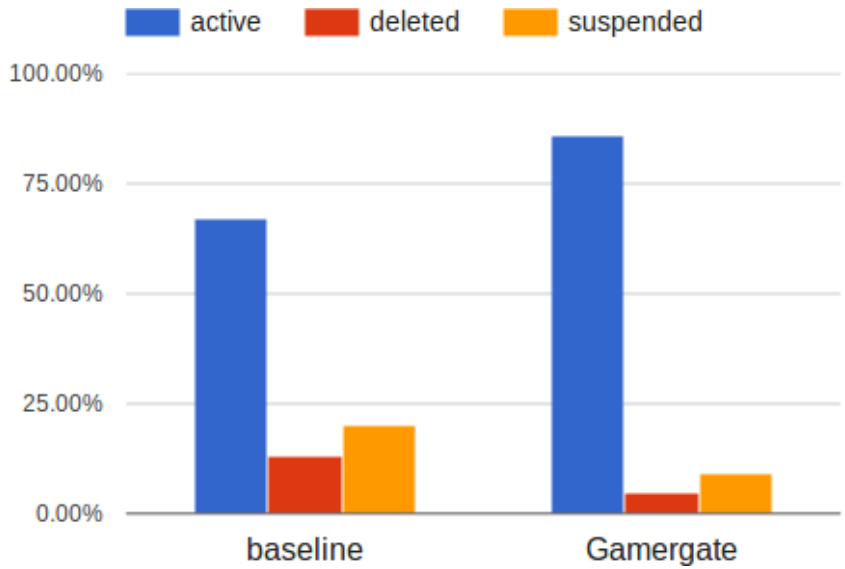


Activity Characteristics - Findings

- GGers tend to have older accounts $->$ greater familiarity with Twitter
- GGers are significantly more active than baseline Twitter users, i.e., more posts
- GGers make more mentions within their posts $->$ higher number of direct attacks compared to random users
- GGers tend to have more friends and followers than random users $->$ the controversy appears to be a clear "us vs. them" situation

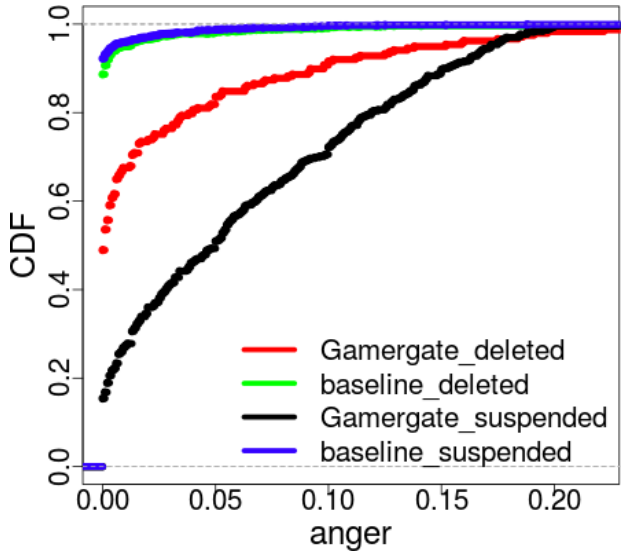
Goal 2

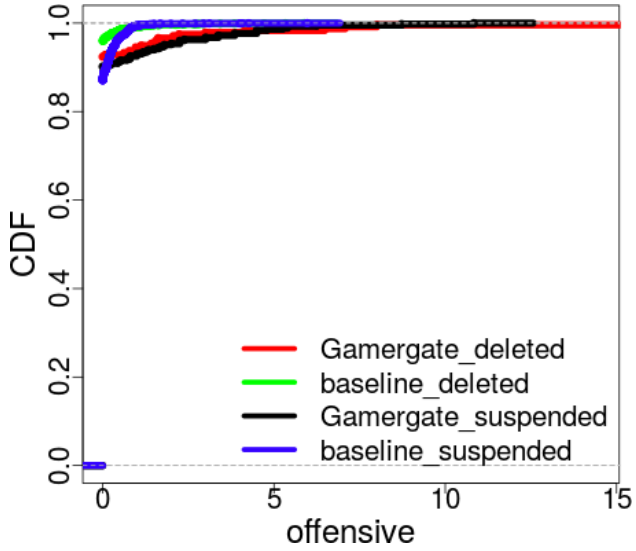
Twitter Reaction to Harassment

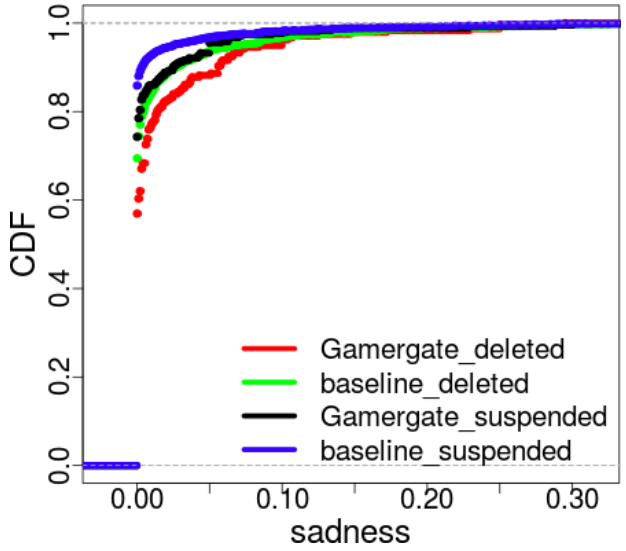


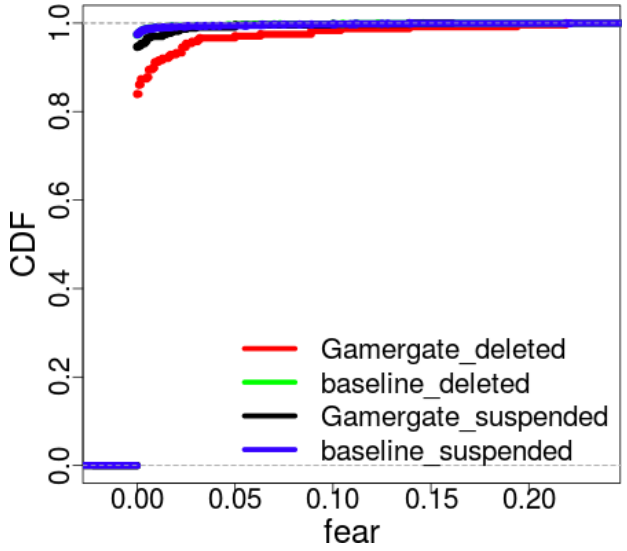
- Focus on a sample of $33k$ users
- Users tend to be suspended more often than delete their accounts
- Random users are more prone to be suspended or delete their accounts than GGers

Emotional Characteristics









Emotional Characteristics - Findings

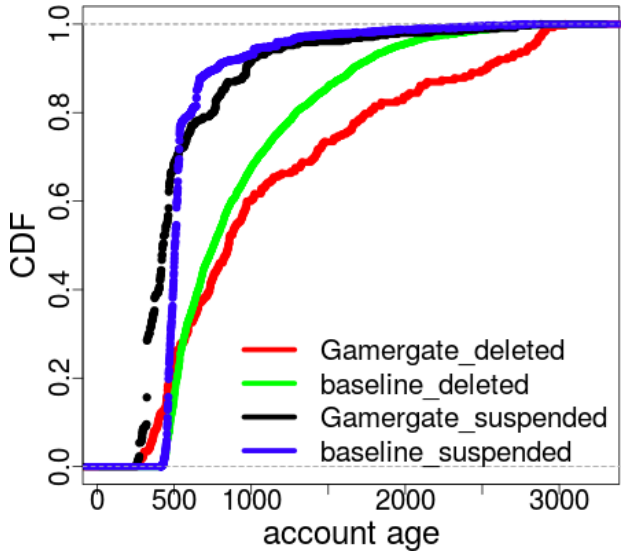
Suspended accounts:

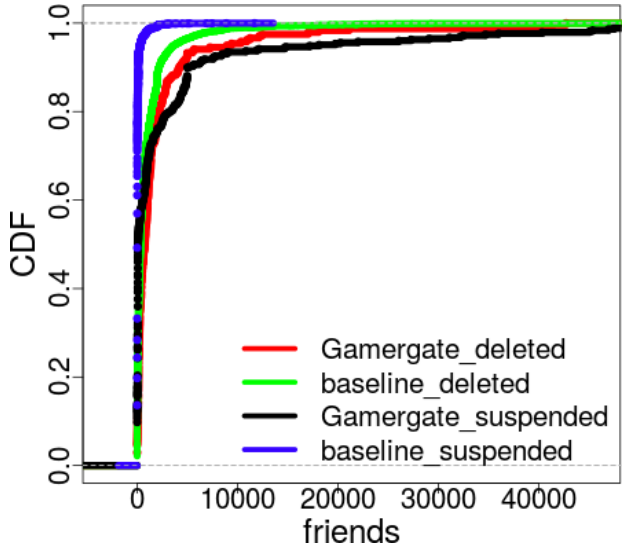
- GGers are expressing more aggressive/repulsive emotions, and offensive language
- 30% of GGers post more negative tweets than baseline users / the rest of the GGers are more positive

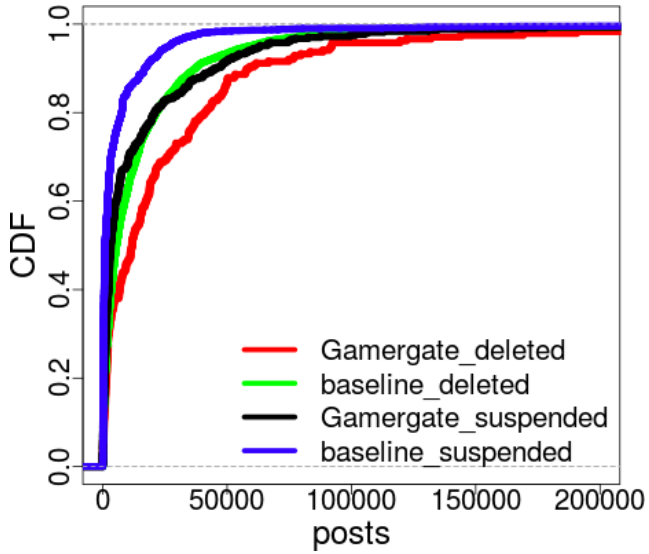
Deleted accounts:

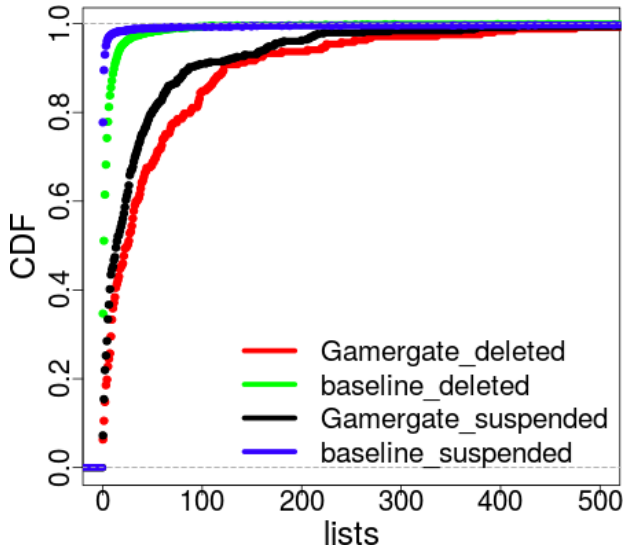
- GGers exhibit higher anger in their posted tweets
- GGers exhibit less joy, but more sadness and fear
- GGers tweet with more negative sentiment
- GGers type less in all uppercase

Activity Characteristics









Activity Characteristics - Findings

- Suspended and deleted GGers are more active overall than baseline users
- Deleted users have less support from their social network (less followers/friends)
- Deleted GGers exhibit the highest activity in comparison to deleted baseline and suspended GGers

Who should be suspended?

- What homogeneity or commonalities users have?
- Group users based on an unsupervised clustering method: k-means.
  - Which is the optimal number of clusters? - EM algorithm.

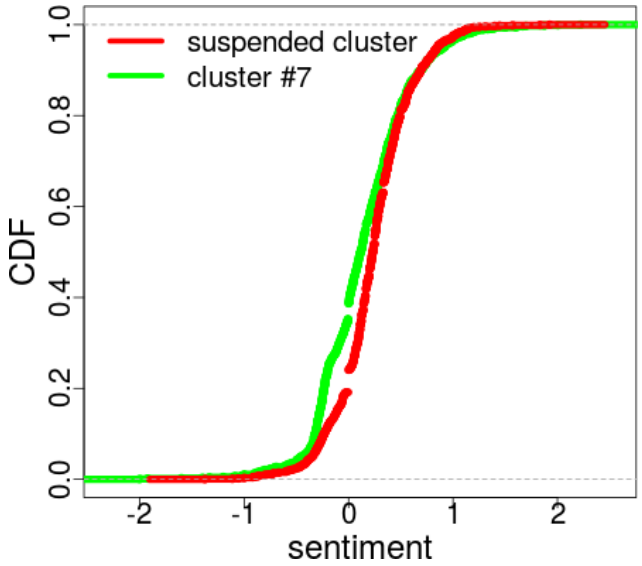
Clustering tendency of baseline users

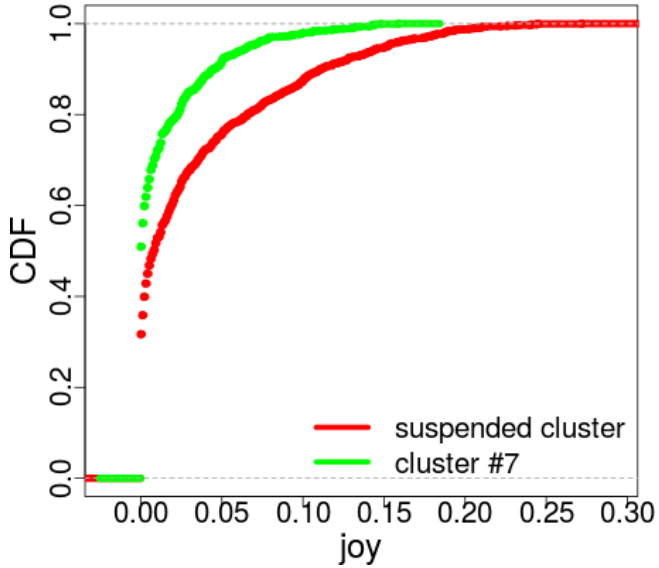
Status →	Cluster	# active	# deleted	# suspended
active	1	4,999	1,501	658
deleted	2	1,984	392	439
suspended	3	4,200	690	3,832
	4	3,333	373	134
	5	1,308	358	120
	6	1,030	169	162
	7	1,525	133	257
	8	433	85	71

Table : Emotional-related features.

Status →	Cluster	# active	# deleted	# suspended
active	1	6,885	1,121	651
deleted	2	882	1,124	63
suspended	3	4,942	574	3,765
	4	1,580	156	74
	5	2,733	594	78
	6	858	51	51
	7	142	24	2
	8	787	57	989

Table : Activity-related features.





Emulating the suspension engine

Classification results based on Gamergate dataset.

	Prec.	Rec.	ROC
active	0.898	0.982	0.747
deleted	0.667	0.008	0.550
suspended	0.669	0.407	0.865
overall (avg.)	0.867	0.886	0.747

Table : Emotional-related features

	Prec.	Rec.	ROC
active	0.937	0.973	0.886
deleted	0.725	0.489	0.804
suspended	0.742	0.591	0.925
overall (avg.)	0.910	0.917	0.886

Table : Activity-related features.

Classification results based on baseline dataset.

	Prec.	Rec.	ROC
active	0.756	0.946	0.742
deleted	0.197	0.022	0.674
suspended	0.803	0.598	0.882
overall (avg.)	0.692	0.755	0.761

Table : Emotional-related features

	Prec.	Rec.	ROC
active	0.806	0.943	0.826
deleted	0.570	0.248	0.806
suspended	0.892	0.718	0.937
overall (avg.)	0.792	0.807	0.846

Table : Activity-related features.

Findings

- GGers were existing Twitter users that were probably drawn to the controversy
- GGers do not exhibit common expressions of online anger
- Suspended GGers tend to become more popular and more active in terms of their posted tweets
- Deleted users exhibit signs of distress, fear, and sadness

Questions

This work has been funded by the European Commission as part of the ENCASE project (H2020-MSCA-RISE), under GA number 691025.