Detecting Aggressors and Bullies on Twitter

WWW 2017

Despoina Chatzakou[1], Nicolas Kourtellis[2], Jeremy Blackburn[2], Emiliano De Cristofaro[3], Gianluca Stringhini[3], Athena Vakali[1]

deppych@csd.auth.gr, nicolas.kourtellis@telefonica.com, jeremy.blackburn@telefonica.com, e.decristofaro@ucl.ac.uk, g.stringhini@ucl.ac.uk, avakali@csd.auth.gr

[1]Aristotle University of Thessaloniki, [2]Telefonica Research, [3]University College London

Abstract

Online social networks constitute an integral part of people's every day social activity.

The existence of **aggressive** and **bullying** phenomena in such spaces is inevitable.

Contributions:

– novel methodology to **collect**, **analyze**, and **label** aggressive and bullying behavior on Twitter

– analysis of bullying and aggressive behavior and extraction of **features** differentiating them from regular users

– **machine learning** approach to automatically detect bullies and aggressors on Twitter

Facts

✓ In 2014, **over 50%** of young people who use social media have reported being cyberbullied.

✓ **Racist** and **sexist** attacks have been reported on Twitter.

✓ The research community has recently focused on detecting **bully** and **aggressive behavior** across various social platforms.

✓ Few works have focused on characterizing **the bullying users** themselves and not only their abusive content.

Definitions

Cyberbullying: repeated and hostile behavior by a group or an individual, using electronic forms of contact.

Cyberaggression: intentional harm delivered by the use of electronic means to a person or a group of people who perceive such acts as offensive, derogatory, harmful, or unwanted.

BULLYING →

Tweet 1 your profile pic sucks! U should wear a mask to hide from the sun 😂 😂 😂 😂

Tweet 2 Our class prom night just got ruined because u showed up. Who invited u anyway?

Tweet 3 Some1 should stalk u and have fun with u..;)

Tweet 4 Don't cry... U can do shit about it...No matter what u do, your pics are out there.:D

Tweet 5 Why do you even show up at school?Nobody cares and neither should u!

Case study: Gamergate

✓ A coordinated campaign of **harassment** in the online world.

✓ It started with a blog post by an ex-boyfriend of independent game developer Zoe Quinn, alleging **sexual improprieties**.

✓ It quickly devolved into a polarizing issue, involving **sexism**, **feminism**, and ``**social justice**,'' taking place on social media like Twitter.

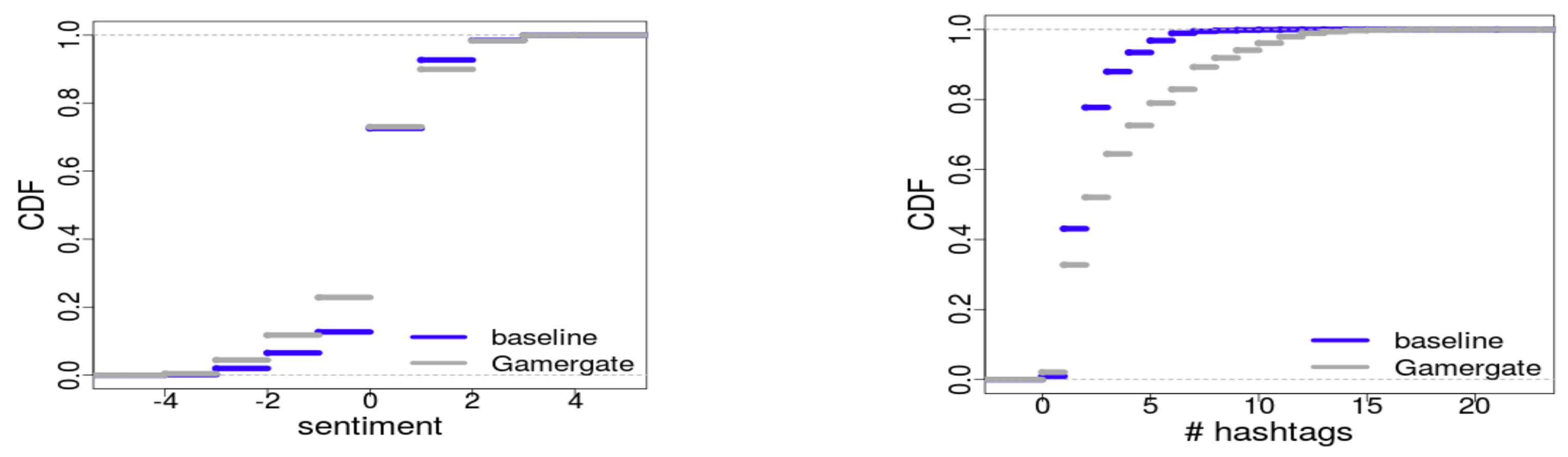
Datasets and Ground truth

The data collection process took place from June to August 2016.

Hate-related: set of 650k tweets based on 309 hashtags associated with bullying and hateful speech.

309 hashtags: **#GamerGate** & 308 hashtags that coexisted within the tweets with the #GamerGate, e.g., #IStandWithHateSpeech, #KillAllNiggers.

Baseline: 1M random tweets.



Ground truth: Crowdsourcing based on the crowdflower.com platform

✓ 1,307 users / 9,484 tweets

✓ 4.5% bully users, 3.4% aggressors, 31.8% spammers, 60.3% normal

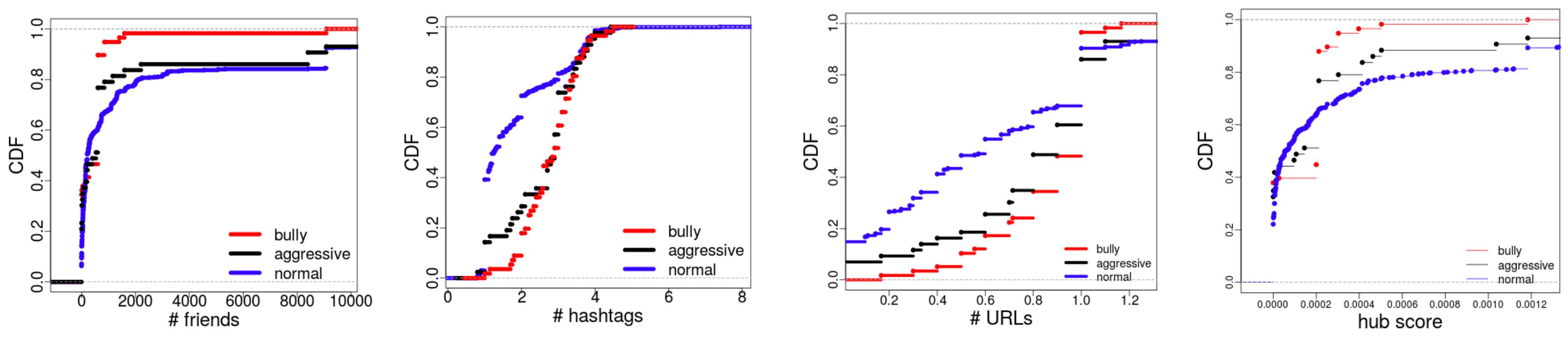
Feature Extraction

Features categorization: **User-based**, **Text-based**, and **Network-based**.

Type	Feature
User (total: 10)	avg. # posts, # days since account creation, verified account # subscribed lists, posts' interarrival time, default profile image? statistics on sessions: total number, avg., median, and STD. of their size
Textual (total: 9)	avg. # hashtags, avg. # emoticons, avg. # upper cases, # URLs avg. sentiment score, avg. emotional scores, hate score avg. word embedding score, avg. curse score
Network (total: 11)	# friends, # followers, hubs, (d=#followers/#friends), authority avg. power diff. with mentioned users, clustering coefficient, reciprocity eigenvector centrality, closeness centrality, louvain modularity

✓ Aggressors and bullies have a propensity to use **more hashtags** within their tweets.

✓ Bullies have **fewer friends** than the other categories.



Information gain: network-based features > user-based > textual ones.

Experimental Results

✓ We experimented with more than **15 machine learning algorithms**.

✓ **Random Forest classifier**: better performance considering both the time for training each classifier and the classification performance.

	Prec.	**Rec.**	**ROC**
bully	0.464	0.448	0.918
aggressive	0.286	0.093	0.868
normal	0.941	0.978	0.925
Avg.	0.878	0.901	0.922

bully	aggres.	normal	
26	7	25	bully (GT)
16	4	23	aggres. (GT)
14	3	770	normal (GT)

Discussion

✓ Various cases are documented where the content of (a set of) posts on online social platforms is **harsh**, **mean**, or **even cruel**.

✓ Detecting the warning signs of cyberbullying poses several difficulties.

✓ We succeed to distinguishing among bullies, aggressors, and typical Twitter user with an average **87.8% precision**, **90.1% recall** and **92.2% AUC**.