January 2020

Mitigating Online Harms at speed and scale

The digital economy urgently needs a new regulatory framework to improve citizens' safety online. This will rebuild public confidence and set clear expectations of companies, allowing our citizens to enjoy more safely the benefits that online services offer.

Illegal and unacceptable content and activity is widespread online, and users are concerned about what they see and experience on the internet. The prevalence of the most serious illegal content and activity, which threatens our national security or the physical safety of children, is unacceptable. Online platforms can be a tool for abuse and bullying, and they can be used to undermine our democratic values and debate. The impact of harmful content and activity can be particularly damaging for children, and there are growing concerns about the potential impact on their mental health and wellbeing (https://www.gov.uk/government/consultations/online-harms-white-paper).

Other online behaviours or content, even if they may not be illegal in all circumstances, can also cause serious harm. The internet can be used to harass, bully or intimidate, especially people in vulnerable groups or in public life. Young adults or children may be exposed to harmful content that relates, for example, to self-harm or suicide. These experiences can have serious psychological and emotional impact. There are also emerging challenges about designed addiction to some digital services and excessive screen time.

In this regard the UK government has set to implement a new regulatory framework to articulate the vision for:

- A free, open and secure internet ensuring freedom of expression online.
- An online environment where companies take effective steps to keep their users safe, and where criminal, terrorist and hostile foreign state activity is not left to contaminate the online space.
- Rules and norms for the internet that discourage harmful behaviour.
- The UK as a thriving digital economy, with a prosperous ecosystem of companies developing innovation in online safety.
- Citizens who understand the risks of online activity, challenge unacceptable behaviours and know how to access help if they experience harm online, with children receiving extra protection.
- A global coalition of countries all taking coordinated steps to keep their citizens safe online.
- Renewed public confidence and trust in online companies and services.

A critical element of the new regulatory framework will be the development of a culture of transparency, trust and accountability. A new statutory duty of care will be established to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services.

For the most serious online offending such as terrorism or child sexual exploitation and abuse (CSE/A), companies are expected to go much further and demonstrate the steps taken to combat the dissemination of associated content and illegal behaviours.

January 2020

Profiling Digital Conversations at scale in near real time

Our view

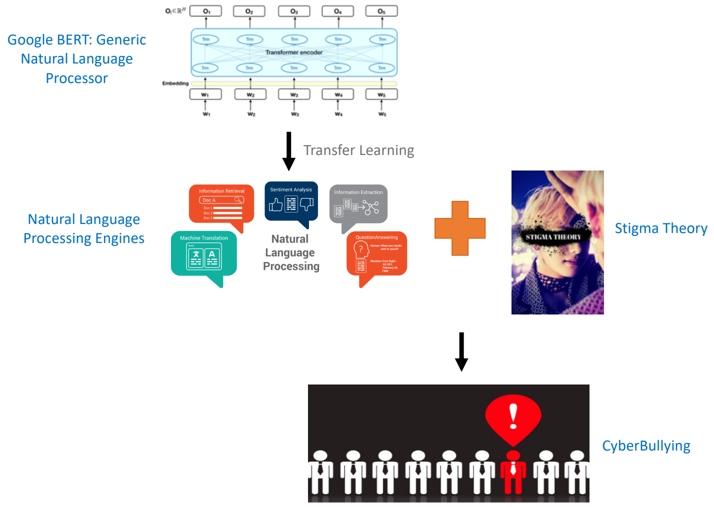
From a regulatory point of view addressing Online Harms should not be any different from addressing abuse in physical spaces: verbal abuse is tolerated neither at home nor at any other public space including schools and workplaces. Moreover, abusers can expect to be prosecuted by the law as a result of their harmful behavior.

In our view differences between Online Harms and Offline Harms are mostly of technical nature whereas in the offline world it is easy to profile abuse behavior (someone being insulted on the street or at school) in the online world the sheer volume of information makes it harder to police and enforce abusive behavior (e.g. profiling child abuse on social media).

Our approach

We believe that feasible solutions to address Online Harms will necessarily combine human agency with AI-powered technologies which would greatly expand the ability to monitor and police the digital world. In this vein we have been busy for the past year developing solutions to help law and security forces the monitoring of conversations taking place in social media.

Our solution relies on the combination of: (1) recent advances in natural language processing to train an engine capable of extracting a set of emotions from human conversations (e.g. tweets) and (2) behavioural theory to infer Online Harms arising from these conversations, refer to the following figure:



From an operational point of view our solution computes, out of a given text, a set of emotions (Anger, Fear, Joy, Love, Optimism, Pessimism, Trust, Toxicity, Threat, Insult, Obscene, Identity_Hate).

Some examples of emotions extracted from real data are:

- Examples of ANGER YES:

Also, since you wrote "starting the production in advance and only launching our campaign towards the tail-end of the manufacturing period", it's perhaps safe to assume that it's already too late to fix the final product, since it has been already produced (almost)? To lessen the change of import duties you should realise we pledge against a reward (a gift). So technically we are not buying anything...

- Examples of FEAR YES:

"*All of us take pride and pleasure in the fact that we are unique, but I'm afraid that when all is said and done the police are right: it all comes down to fingerprints."\n—*

- Examples of JOY YES:

Yea, so happy for you Dad's!! In memory of Holly, in celebration of her life. Always in the hearts of her family and friends.

- Examples of INSULT YES:

@GretaThunberg Greta Disgusting Brat, you're a Scam a Liar Fraud, as Scoundrel as Macron ...

The ability to compute negative emotions (Toxicity, Insult, Obscene, Threat, Identity_Hate) in near real time at scale enables digital companies to profile Online Harming and act pre-emptively before it spreads and causes further damage.

Competitive advantage

Our solution is: (1) highly adaptive to new languages and specific contexts (jargon, teenagers' language) thanks to transfer learning techniques (REF gogole bert), (2) relies on state-of-the-art technologies to achieve outstanding levels of predictive accuracy.

As of January 2020 our technologies exhibit an outstanding level of predictive accuracy, refer to the following table:

Anger: {'auc': 0.9686095, 'eval_accuracy': 0.9696226, 'f1_score': 0.95550555, 'false_negatives': 38.0, 'false_positives': 61.0, 'loss': 0.108323365, 'precision': 0.94572955, 'recall': 0.96548593, 'true_negatives': 2097.0, 'true_positives': 1063.0,	**Anticipation:** {'auc': 0.60829467, 'eval_accuracy': 0.89444613, 'f1_score': 0.35338342, 'false_negatives': 331.0, 'false_positives': 13.0, 'loss': 0.3266065, 'precision': 0.8785047, 'recall': 0.22117648, 'true_negatives': 2821.0, 'true_positives': 94.0,	**Disgust:** {'auc': 0.49833164, 'eval_accuracy': 0.65879107, 'f1_score': 0.5043598, 'false_negatives': 1093.0, 'false_positives': 19.0, 'loss': 0.6443392, 'precision': 0.24, 'recall': 0.005459509, 'true_negatives': 2141.0, 'true_positives': 6.0,
Fear: {'auc': 0.5, 'eval_accuracy': 0.8511813, 'f1_score': 0.25908118, 'false_negatives': 485.0, 'false_positives': 0.0, 'loss': 0.42383027, 'precision': 0.0, 'recall': 0.0, 'true_negatives': 2774.0, 'true_positives': 0.0,	**Joy.** {'auc': 0.9894273, 'eval_accuracy': 0.98956734, 'f1_score': 0.98821074, 'false_negatives': 17.0, 'false_positives': 17.0, 'loss': 0.045616243, 'precision': 0.9882108, 'recall': 0.9882108, 'true_negatives': 1800.0, 'true_positives': 1425.0,	**Love:** {'auc': 0.94728124, 'eval_accuracy': 0.97084993, 'f1_score': 0.9083895, 'false_negatives': 45.0, 'false_positives': 50.0, 'loss': 0.099055305, 'precision': 0.9040307, 'recall': 0.9127907, 'true_negatives': 2693.0, 'true_positives': 471.0,

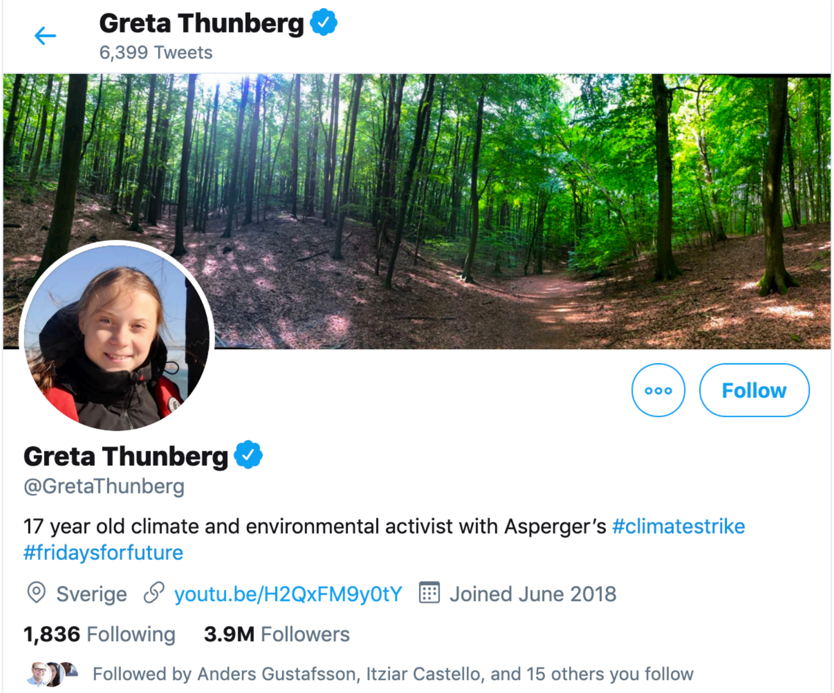
Optimism: {'auc': 0.926086, 'eval_accuracy': 0.9254373, 'f1_score': 0.89725155, 'false_negatives': 82.0, 'false_positives': 161.0, 'loss': 0.20371501, 'precision': 0.86824876, 'recall': 0.92825896, 'true_negatives': 1955.0, 'true_positives': 1061.0,	**Pessimism:** {'auc': 0.9051725, 'eval_accuracy': 0.97177047, 'f1_score': 0.86968833, 'false_negatives': 68.0, 'false_positives': 24.0, 'loss': 0.11386829, 'precision': 0.92749244, 'recall': 0.81866664, 'true_negatives': 2860.0, 'true_positives': 307.0,	**Sadness:** {'auc': 0.5, 'eval_accuracy': 0.7054311, 'f1_score': 0.45508412, 'false_negatives': 960.0, 'false_positives': 0.0, 'loss': 0.6136402, 'precision': 0.0, 'recall': 0.0, 'true_negatives': 2299.0, 'true_positives': 0.0,

Our technologies run on the cloud (linux-based infrastructure, multi-GPU) and achieve a processing performance of 1000 texts per minute per server and 12 different emotions. Our solution is completely modular and scales linearly with the number of servers running in parallel in the cloud.

Following sections will outline the results of two use cases related to Online Harms in public life and cyberbullying.

January 2020

Tackling Cyberbullying: the case of Greta Thunberg



Greta Thunberg
6,399 Tweets

Greta Thunberg
@GretaThunberg

17 year old climate and environmental activist with Asperger's #climatestrike #fridaysforfuture

Sverige youtu.be/H2QxFM9y0tY Joined June 2018

1,836 Following **3.9M** Followers

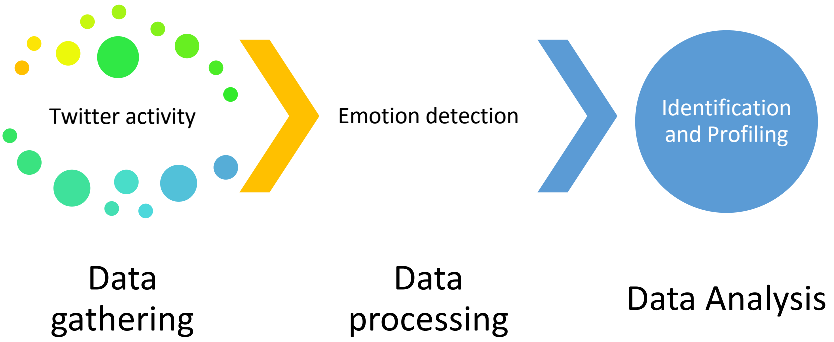
Followed by Anders Gustafsson, Itziar Castello, and 15 others you follow

Greta Thunberg, a seventeen-year-old teenager has risen as a prominent public figure spearheading the climate change movement worldwide. Greta Thunberg is quite popular on social media with 3.9 million followers.

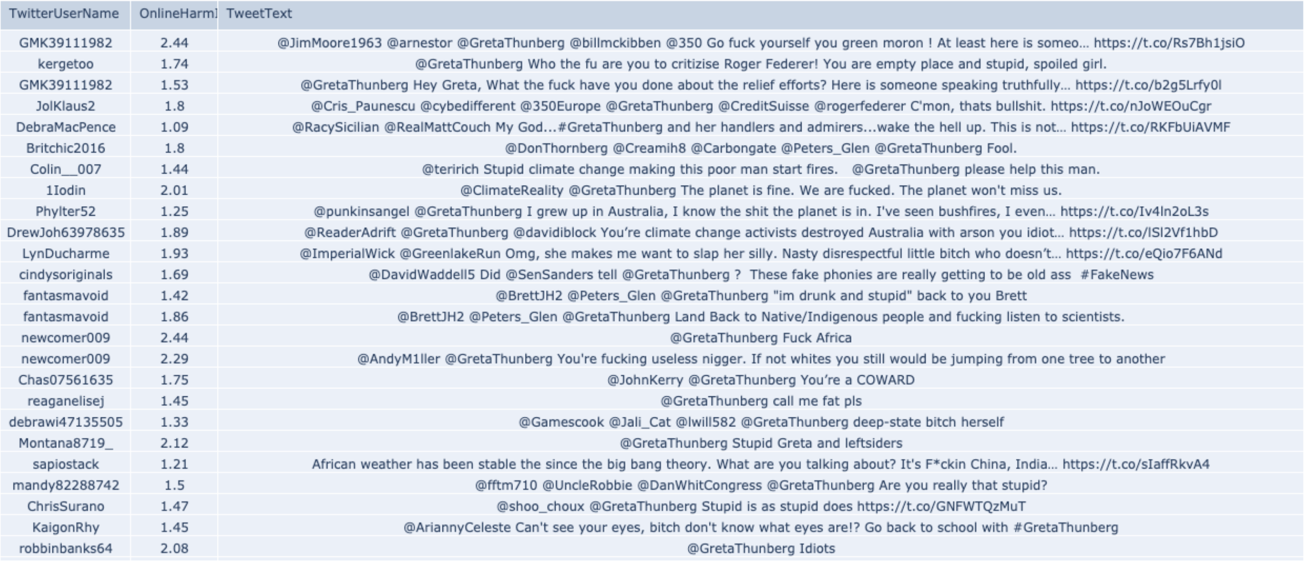
Unfortunately, Greta has been subject to unacceptable abuse on social media platforms even by prominent citizens: *https://www.theguardian.com/commentisfree/2019/dec/14/trump-president-greta-thunberg-bullying*

This unfortunate situation led us to develop a use case testing our solution to address cyberbullying aimed at Greta.

We compiled a sample of 20K tweets referring to Greta Thunberg either by her twitter handle, @GretaThunberg, or by the hashtag #gretathunberg. We processed each message using our emotion detection engine to compute levels of insult, threat, toxicity, severe_toxicity, obscenity and identity_hate. Finally, we applied conventional statistical techniques to profile and identify online abusers. Refer to the following figure.

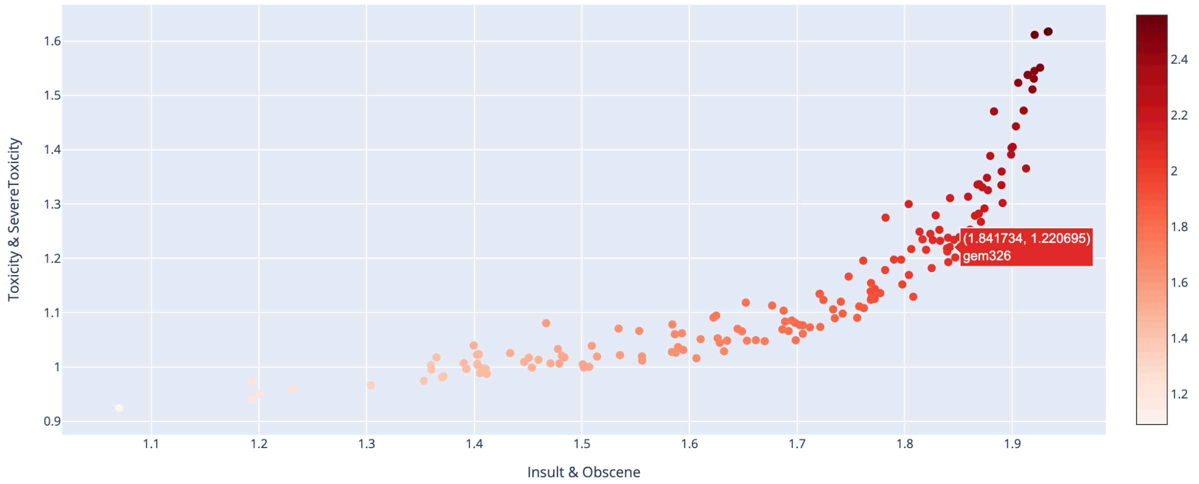


The results of our analysis reveal that, for the sample considered, 1% of the messages could be considered abusive and harmful, refer to the following table (also available online at: *https://chart-studio.plot.ly/~davidlopezberzosa/442*)



TwitterUserName	OnlineHarm	TweetText
GMK39111982	2.44	@JimMoore1963 @arnestor @GretaThunberg @billmckibben @350 Go fuck yourself you green moron ! At least here is someo… https://t.co/Rs7Bh1jsiO
kergetoo	1.74	@GretaThunberg Who the fu are you to critizise Roger Federer! You are empty place and stupid, spoiled girl.
GMK39111982	1.53	@GretaThunberg Hey Greta, What the fuck have you done about the relief efforts? Here is someone speaking truthfully… https://t.co/b2g5Lrfy0l
JolKlaus2	1.8	@Cris_Paunescu @cybedifferent @350Europe @GretaThunberg @CreditSuisse @rogerfederer C'mon, thats bullshit. https://t.co/nJoWEOuCgr
DebraMacPence	1.09	@RacySicilian @RealMattCouch My God…#GretaThunberg and her handlers and admirers…wake the hell up. This is not… https://t.co/RKFbUiAVMF
Britchic2016	1.8	@DonThornberg @Creamih8 @Carbongate @Peters_Glen @GretaThunberg Fool.
Colin__007	1.44	@teririch Stupid climate change making this poor man start fires. @GretaThunberg please help this man.
1Iodin	2.01	@ClimateReality @GretaThunberg The planet is fine. We are fucked. The planet won't miss us.
Phylter52	1.25	@punkinsangel @GretaThunberg I grew up in Australia, I know the shit the planet is in. I've seen bushfires, I even… https://t.co/Iv4ln2oL3s
DrewJoh63978635	1.89	@ReaderAdrift @GretaThunberg @davidiblock You're climate change activists destroyed Australia with arson you idiot… https://t.co/lSl2Vf1hbD
LynDucharme	1.93	@ImperialWick @GreenlakeRun Omg, she makes me want to slap her silly. Nasty disrespectful little bitch who doesn't… https://t.co/eQio7F6ANd
cindysoriginals	1.69	@DavidWaddell5 Did @SenSanders tell @GretaThunberg ? These fake phonies are really getting to be old ass #FakeNews
fantasmavoid	1.42	@BrettJH2 @Peters_Glen @GretaThunberg "im drunk and stupid" back to you Brett
fantasmavoid	1.86	@BrettJH2 @Peters_Glen @GretaThunberg Land Back to Native/Indigenous people and fucking listen to scientists.
newcomer009	2.44	@GretaThunberg Fuck Africa
newcomer009	2.29	@AndyM1ller @GretaThunberg You're fucking useless nigger. If not whites you still would be jumping from one tree to another
Chas07561635	1.75	@JohnKerry @GretaThunberg You're a COWARD
reaganelisej	1.45	@GretaThunberg call me fat pls
debrawi47135505	1.33	@Gamescook @Jali_Cat @lwill582 @GretaThunberg deep-state bitch herself
Montana8719_	2.12	@GretaThunberg Stupid Greta and leftsiders
sapiostack	1.21	African weather has been stable the since the big bang theory. What are you talking about? It's F*ckin China, India… https://t.co/sIaffRkvA4
mandy82288742	1.5	@fftm710 @UncleRobbie @DanWhitCongress @GretaThunberg Are you really that stupid?
ChrisSurano	1.47	@shoo_choux @GretaThunberg Stupid is as stupid does https://t.co/GNFWTQzMuT
KaigonRhy	1.45	@AriannyCeleste Can't see your eyes, bitch don't know what eyes are!? Go back to school with #GretaThunberg
robbinbanks64	2.08	@GretaThunberg Idiots

Our solution achieved a processing performance of 1000 messages per minute (single GPU server). Further performance can be achieved by deploying additional servers running in parallel.
We were able to identify not only harmful content but also the originators of the abuse, refer to the following figure (also available online at: *https://plot.ly/~davidlopezberzosa/438/)*



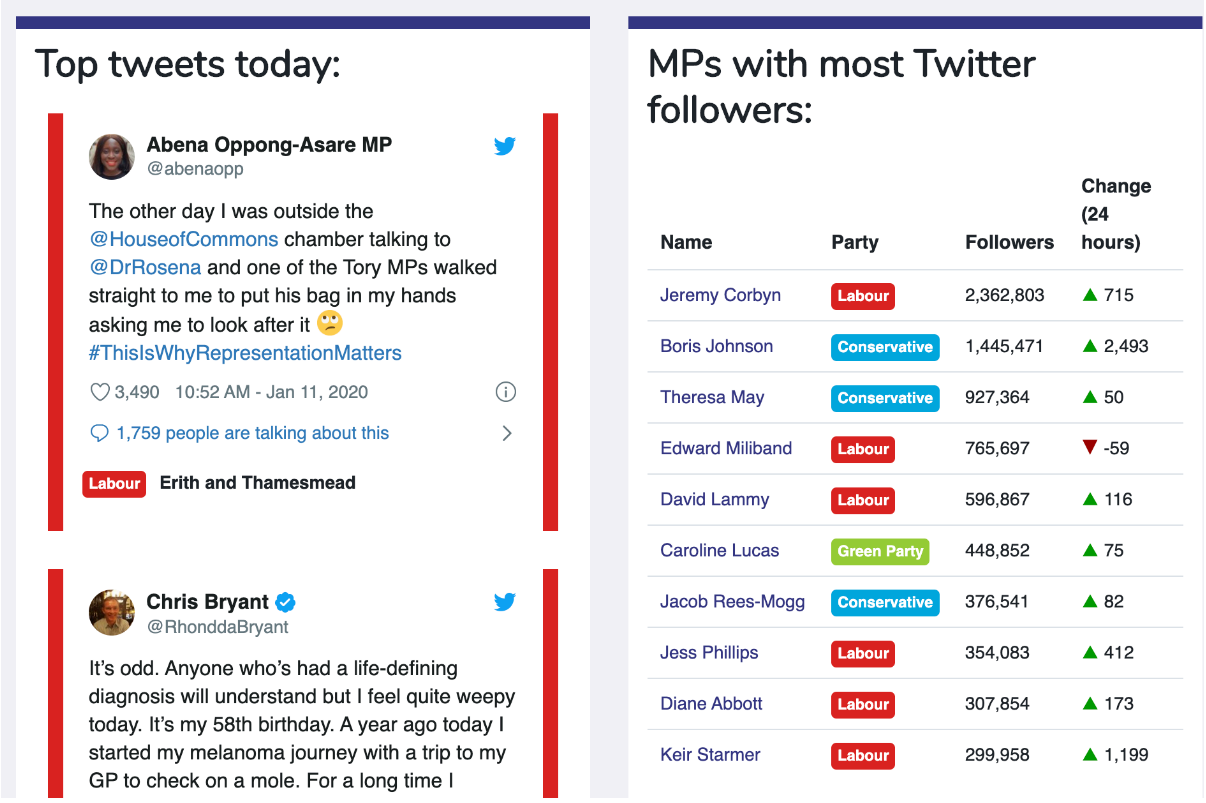
Implications for digital companies and regulators.

The use case of Greta Thunberg exposes the urgency and relevance of protecting citizens online. The results of our analysis provide arguments supporting the use of Artificial Intelligence to monitor abusive behaviour online.

We believe AI-powered technologies can support digital companies to fulfil their duty of care of online citizens in a cost-effective manner.

January 2020

Tackling Cyberbullying:
Toxicity in UK Politics



Top tweets today:

Abena Oppong-Asare MP
@abenaopp

The other day I was outside the @HouseofCommons chamber talking to @DrRosena and one of the Tory MPs walked straight to me to put his bag in my hands asking me to look after it 🙁 #ThisIsWhyRepresentationMatters

♡ 3,490 10:52 AM - Jan 11, 2020

1,759 people are talking about this

Labour **Erith and Thamesmead**

Chris Bryant
@RhonddaBryant

It's odd. Anyone who's had a life-defining diagnosis will understand but I feel quite weepy today. It's my 58th birthday. A year ago today I started my melanoma journey with a trip to my GP to check on a mole. For a long time I

MPs with most Twitter followers:

Name	Party	Followers	Change (24 hours)
Jeremy Corbyn	Labour	2,362,803	▲ 715
Boris Johnson	Conservative	1,445,471	▲ 2,493
Theresa May	Conservative	927,364	▲ 50
Edward Miliband	Labour	765,697	▼ -59
David Lammy	Labour	596,867	▲ 116
Caroline Lucas	Green Party	448,852	▲ 75
Jacob Rees-Mogg	Conservative	376,541	▲ 82
Jess Phillips	Labour	354,083	▲ 412
Diane Abbott	Labour	307,854	▲ 173
Keir Starmer	Labour	299,958	▲ 1,199

Building on the experience gained from the previous use case we decided to further test and validate our solution in another context: online abuse against public figures.

We decided to further test and validate our solution in the context of a use case specific to candidates for the House of Parliament in the recent UK election.

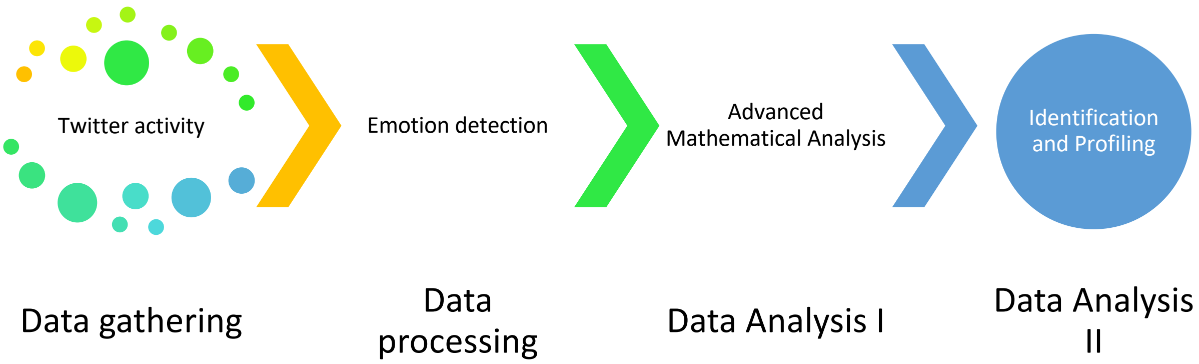
Politicians, as public figures, are unfortunately exposed to all sorts of verbal abuse both offline and online with resulting Online Harms:

https://www.nytimes.com/2019/11/01/world/europe/women-parliament-abuse.html

This case is relevant not only given the importance of the collective under consideration but also because from a technical point of view it entails higher levels of complexity. If in the previous case the target was known in advance (Greta Thunberg) in the present use case we consider many targets (all candidates to HoP) and many potential online abusers (all candidates to HoP).

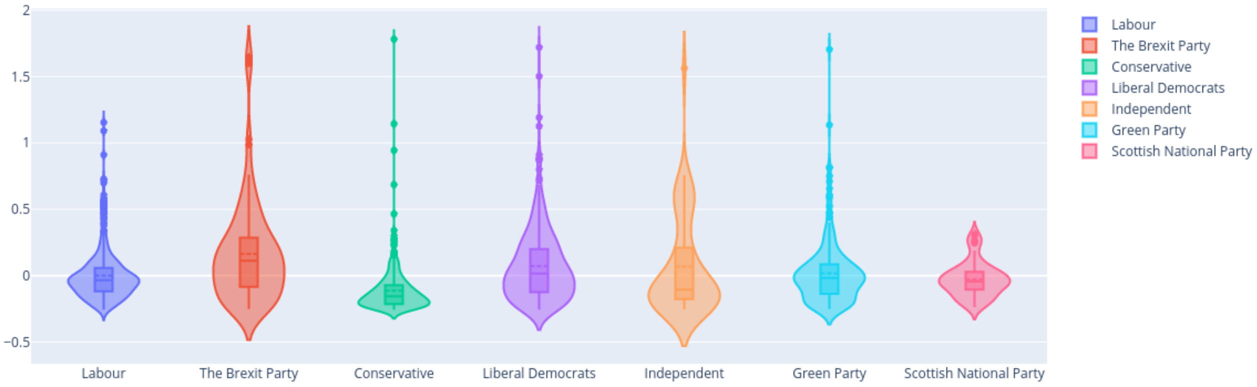
Whereas the complexity of Greta Thunberg lied just in the computation of emotions in this use case there is an additional layer of complexity in finding all the combinations of sources and recipients of the online abuse.

We extracted all tweets from all candidates to the UK House of Parliament (a total of 197476 tweets from 1953 candidates).We processed each message using our emotion detection engine to compute levels of insult, threat, toxicity, severe_toxicity, obscenity and identity_hate. We applied graph theory to compute all the relationships established through the association of twitter handles with twitter messages (i.e. whenever someone wrote a tweet referring to a candidate).
Finally, we applied conventional statistical techniques to profile and identify online abusers. Refer to the following figure.

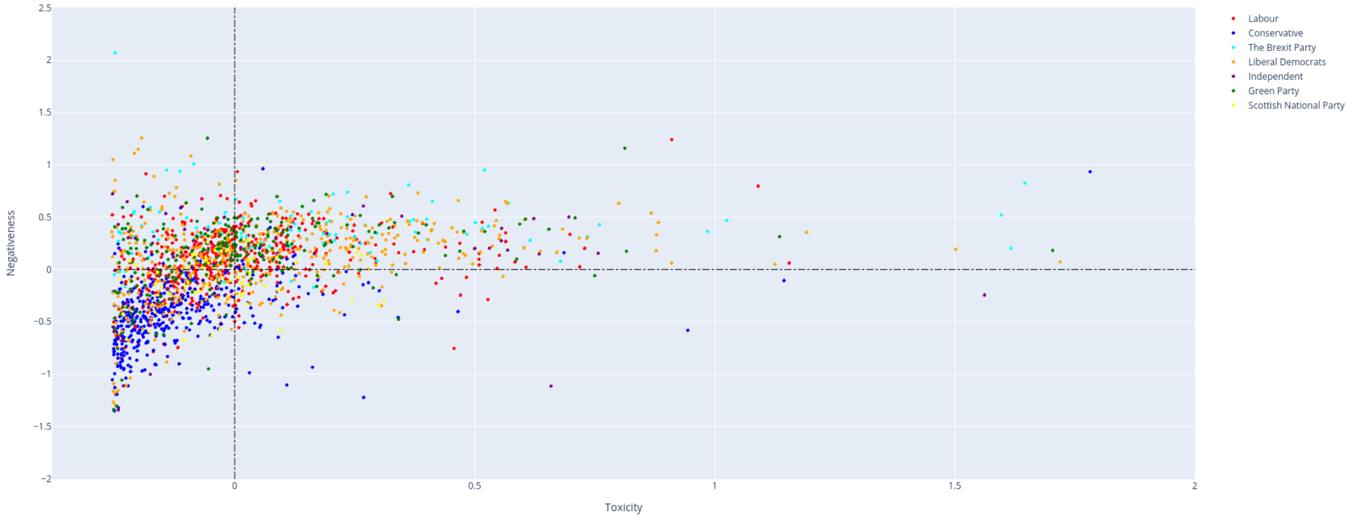


The following figure summarizes the levels of verbal abuse (measured in terms of toxic language) between political parties (also available online at: *https://plot.ly/~davidlopezberzosa/412/*).

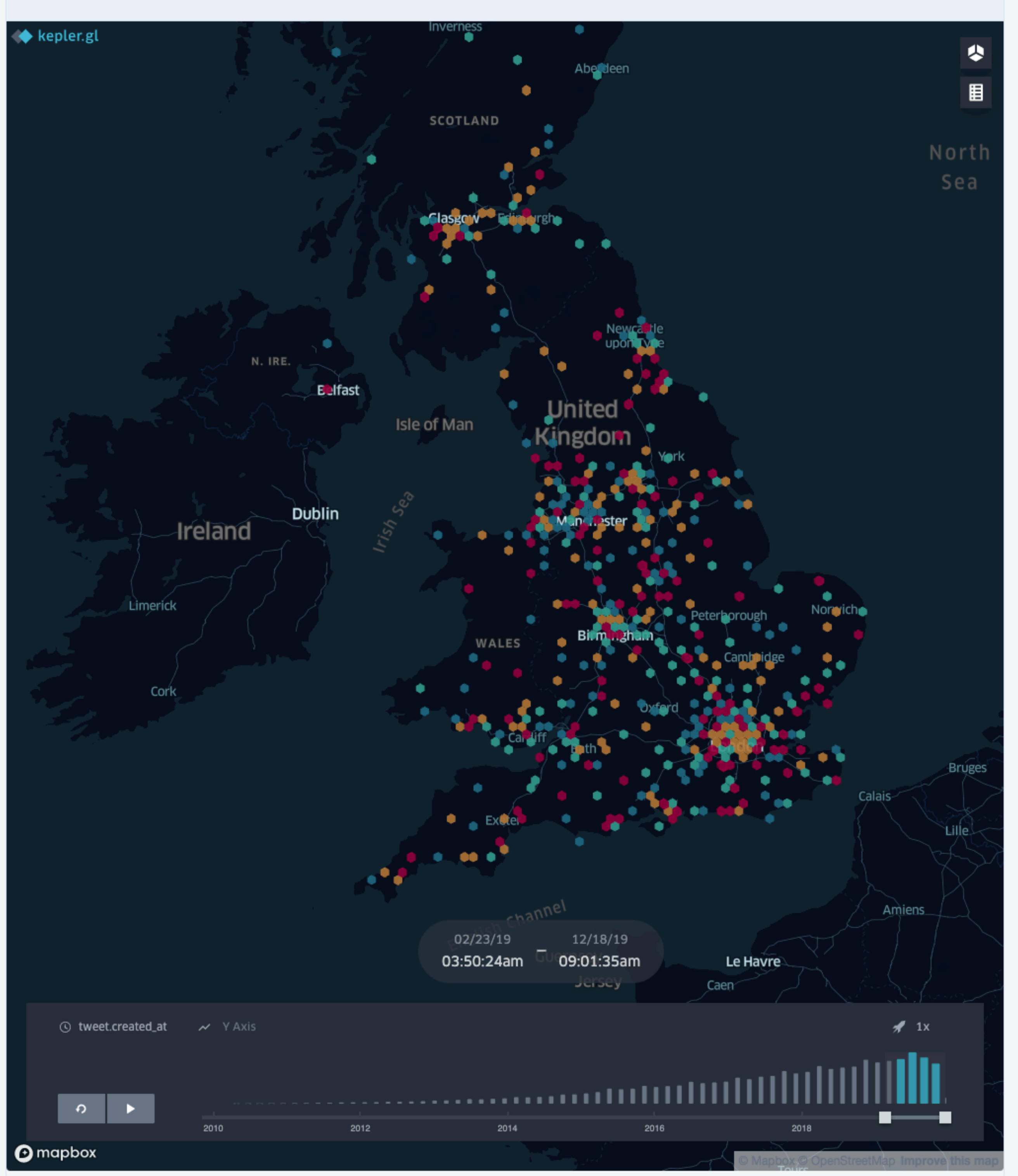
We can conclude that overall political parties are civilized when referring to other candidates. The Brexit Party and Liberal Democrats are above average in terms of toxic comments made on twitter.



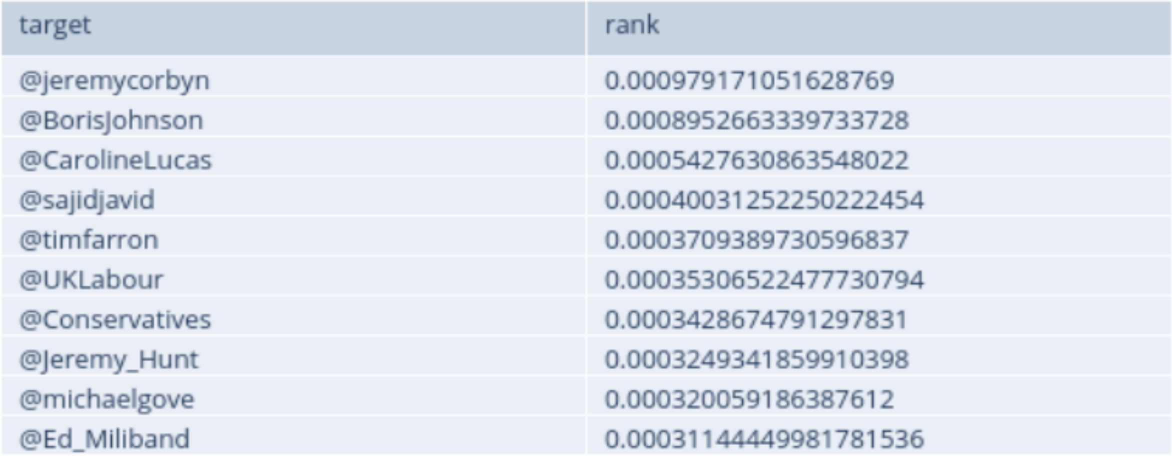
The following figure situates each candidate according to his/her levels of toxicity and negativeness when referring to other candidates on twitter (also available at: *https://plot.ly/~davidlopezberzosa/416/*). The results confirm that most politicians are respectful of others with few exceptions (dots situated at the right-hand side of the figure).



The following figure provides a general overview of toxicity levels across constituencies during the election period (available at: https://chart-studio.plot.ly/dashboard/davidlopezberzosa:430/view). We observe above average levels of toxicity in some constituencies in large cities.

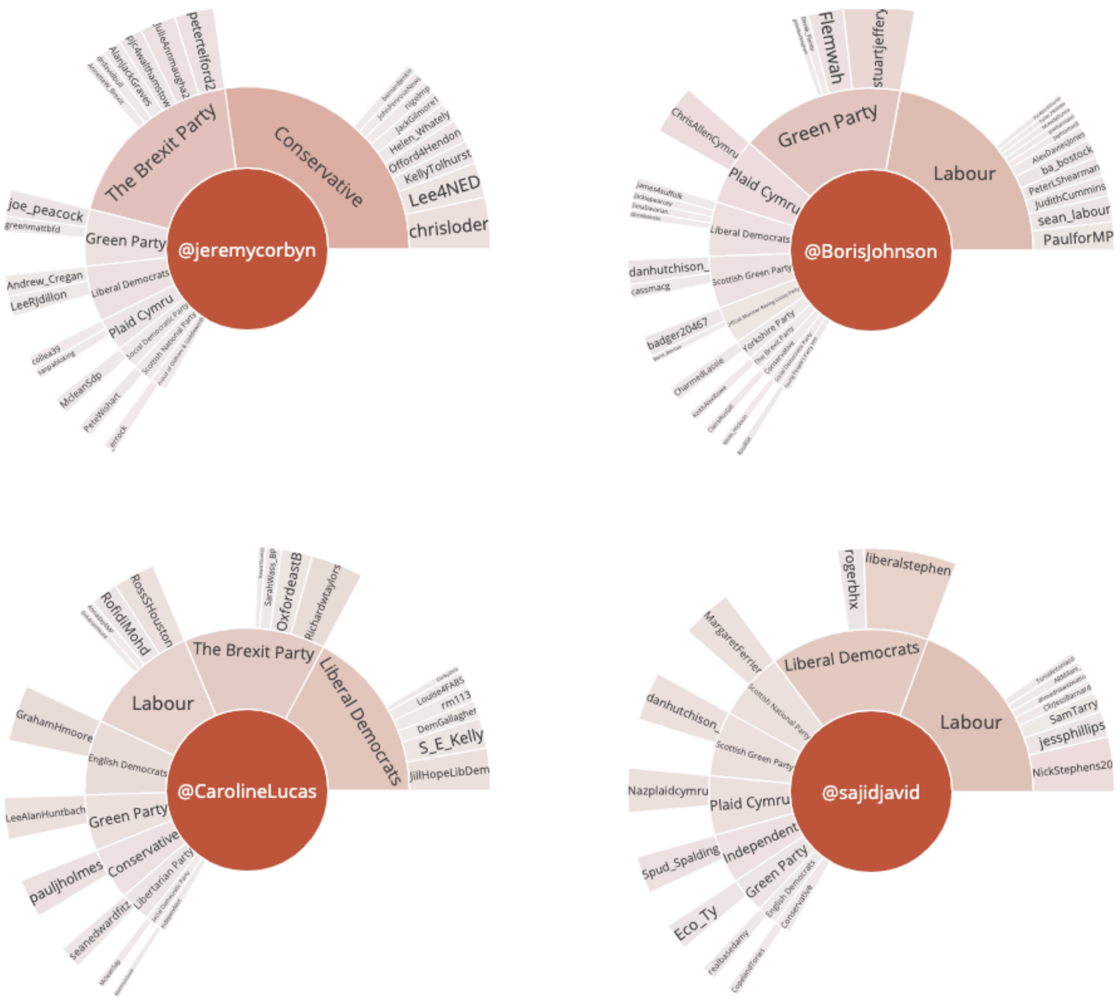


The following table lists politicians exposed to above average levels of toxicity:

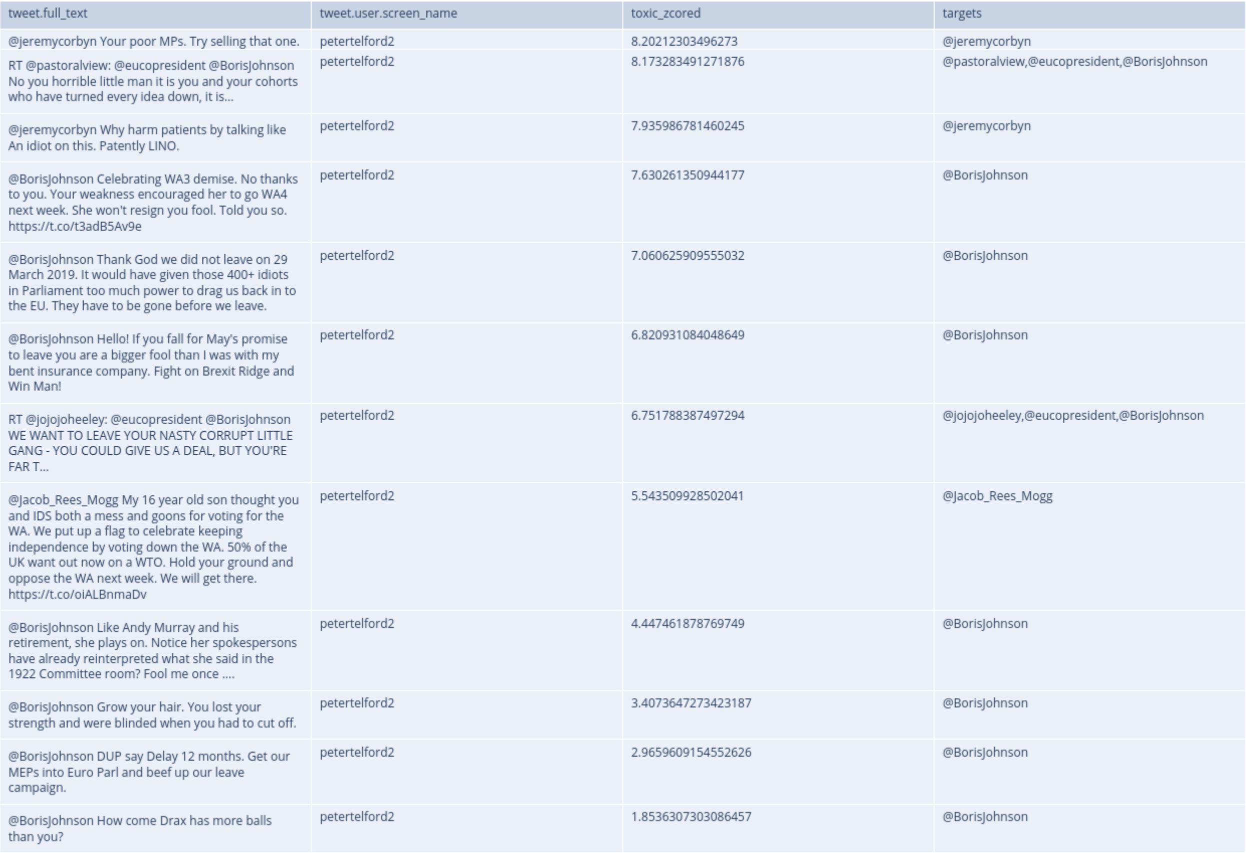


target	rank
@jeremycorbyn	0.000979171051628769
@BorisJohnson	0.0008952663339733728
@CarolineLucas	0.0005427630863548022
@sajidjavid	0.00040031252250222454
@timfarron	0.0003709389730596837
@UKLabour	0.00035306522477730794
@Conservatives	0.0003428674791297831
@Jeremy_Hunt	0.0003249341859910398
@michaelgove	0.000320059186387612
@Ed_Miliband	0.00031144449981781536

The following figure provides a visual summary of which politicians exhibit above average levels of toxicity against Boris Johnson, Caroline Lucas, Jeremy Corbyn and Sajid Javid (available at: https://plot.ly/~davidlopezberzosa/424/).



The following table provides evidence of toxic comments made by candidate Peter Telford (Brexit Party) against other candidates (also available at: https://plot.ly/~davidlopezberzosa/436/)



tweet.full_text	tweet.user.screen_name	toxic_zcored	targets
@jeremycorbyn Your poor MPs. Try selling that one.	petertelford2	8.20212303496273	@jeremycorbyn
RT @pastoralview: @eucopresident @BorisJohnson No you horrible little man it is you and your cohorts who have turned every idea down, it is…	petertelford2	8.173283491271876	@pastoralview,@eucopresident,@BorisJohnson
@jeremycorbyn Why harm patients by talking like An idiot on this. Patently LINO.	petertelford2	7.935986781460245	@jeremycorbyn
@BorisJohnson Celebrating WA3 demise. No thanks to you. Your weakness encouraged her to go WA4 next week. She won't resign you fool. Told you so. https://t.co/t3adB5Av9e	petertelford2	7.630261350944177	@BorisJohnson
@BorisJohnson Thank God we did not leave on 29 March 2019. It would have given those 400+ idiots in Parliament too much power to drag us back in to the EU. They have to be gone before we leave.	petertelford2	7.060625909555032	@BorisJohnson
@BorisJohnson Hello! If you fall for May's promise to leave you are a bigger fool than I was with my bent insurance company. Fight on Brexit Ridge and Win Man!	petertelford2	6.820931084048649	@BorisJohnson
RT @jojojoheeley: @eucopresident @BorisJohnson WE WANT TO LEAVE YOUR NASTY CORRUPT LITTLE GANG - YOU COULD GIVE US A DEAL, BUT YOU'RE FAR T…	petertelford2	6.751788387497294	@jojojoheeley,@eucopresident,@BorisJohnson
@Jacob_Rees_Mogg My 16 year old son thought you and IDS both a mess and goons for voting for the WA. We put up a flag to celebrate keeping independence by voting down the WA. 50% of the UK want out now on a WTO. Hold your ground and oppose the WA next week. We will get there. https://t.co/oiALBnmaDv	petertelford2	5.543509928502041	@Jacob_Rees_Mogg
@BorisJohnson Like Andy Murray and his retirement, she plays on. Notice her spokespersons have already reinterpreted what she said in the 1922 Committee room? Fool me once ….	petertelford2	4.447461878769749	@BorisJohnson
@BorisJohnson Grow your hair. You lost your strength and were blinded when you had to cut off.	petertelford2	3.4073647273423187	@BorisJohnson
@BorisJohnson DUP say Delay 12 months. Get our MEPs into Euro Parl and beef up our leave campaign.	petertelford2	2.9659609154552626	@BorisJohnson
@BorisJohnson How come Drax has more balls than you?	petertelford2	1.8536307303086457	@BorisJohnson

Acknowledgments

We gratefully acknowledge the support of the Alan Turing Institute, University of Exeter, INCIBE (Ventures Programme), Google (Start-up acceleration programme) and Amazon (cloud computing infrastructure).