Assessing Toxicity in Wikipedia Comments

Jonathan Innis & Gabriel Britain

Disclaimer: Some comments in this presentation may be offensive to certain viewers. The comments in this presentation do not reflect the opinions of the creators/presenters and are used purely for academic purposes.

Purpose

- Identifying toxicity can prevent users from abusing communication platforms
- Much more efficient than review by human moderators
- Most comments are posted at early hours of the morning (3am) and will be uncaught by human moderators for hours

How to Deal With a $759 Million Lottery Jackpot

Toxic

Carl Hollis · 2 hours ago

two ignorant idiots you two are.

↑0 ↓0 Edit Reply

Approve Spam Delete

Arpaio Pardon Would Show Contempt for Constitution

OnKilter · 5 hours ago

STFU racist pig.
Just STFU.

↑0 ↓0 Edit Reply

Approve Spam Delete

92% similar to comments people said were "toxic" SEEM WRONG?

Nieman Lab is a great website — only an idiot like you would think some other website could possibly be better. You dumb jerk.

[redacted] Nov 30, 2014 +2

First of all, A for effort! But I wasn't a racist [redacted] like you were, so my grammar is irrelevant (so I'm not a hypocrite, although that's a big word, you should be proud). Also, I should point out that yours didn't improve, so we got nowhere with you. Your spelling makes me inclined to think you're a 'dirty [redacted] yourself! And I hope at the end that you weren't threatening to kill me. I'll forgive you because you seem cranky, so I'd suggest a nap, you mouth-breathing, stagnant cesspool of human trash.

Show less

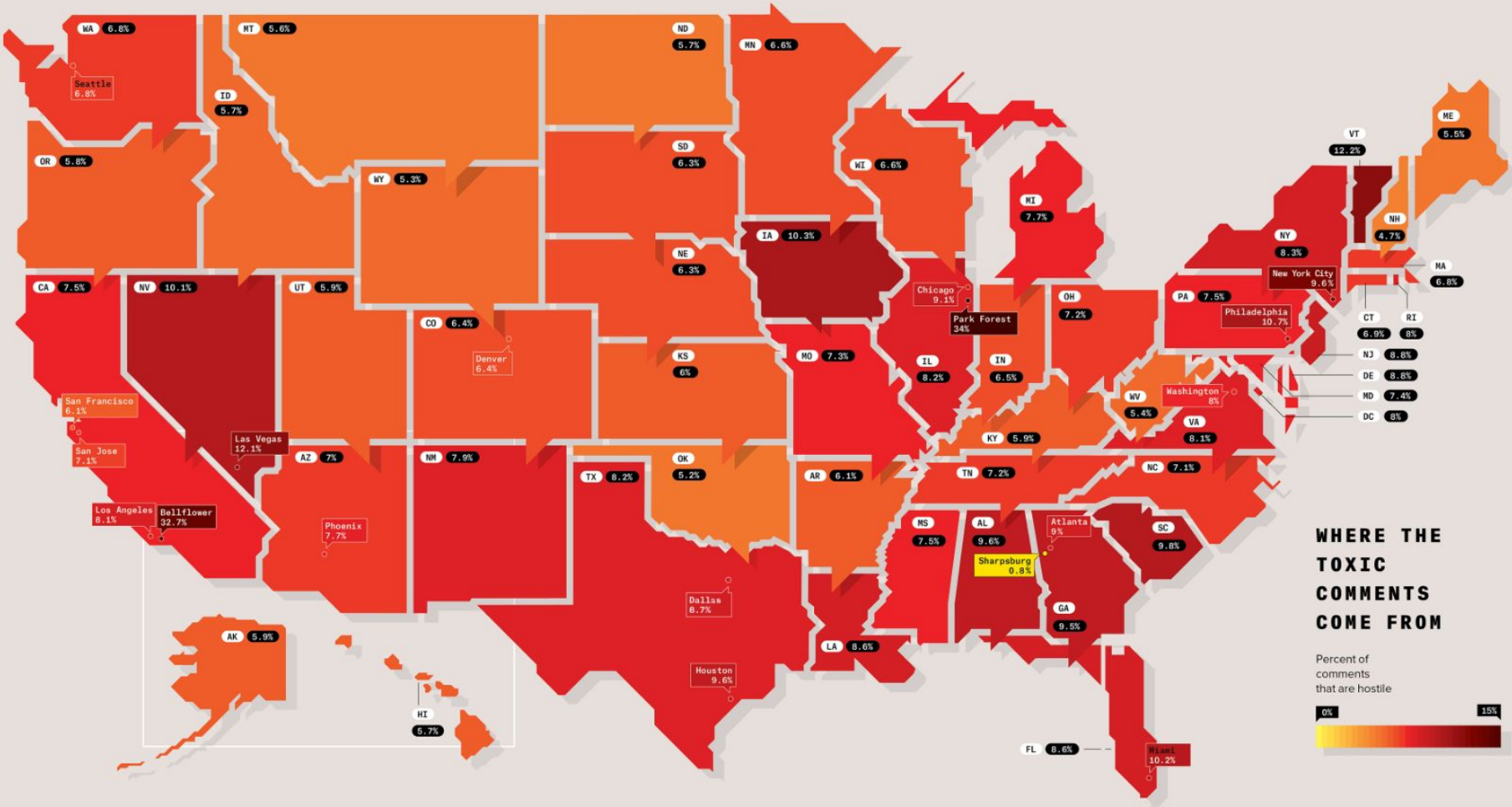
98% similar to comments people said were "toxic" SEEM WI

You're a stupid idiot!

UnknownArchive 1 week ago

#1 [redacted] this feminist [redacted] and all the damage she did to thatguywiththeglasses.
#2 see #1

Reply · 15



https://www.wired.com/2017/08/internet-trol
/

Dataset Source

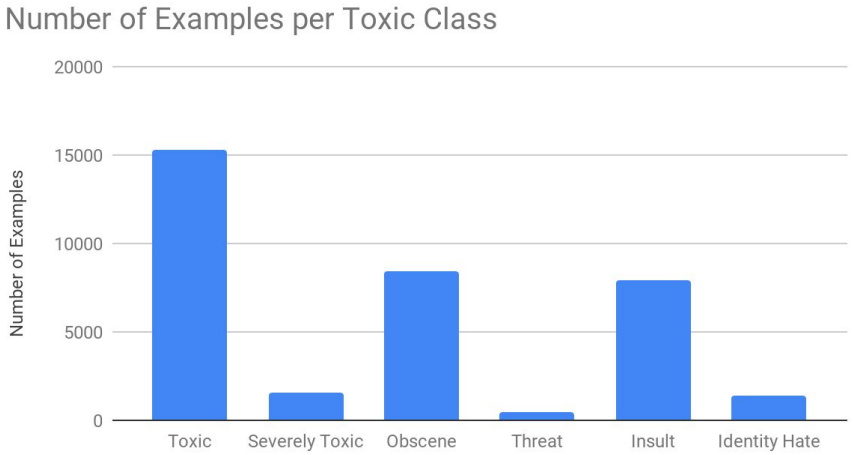
id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. T	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is con	0	0	0	0	0	0
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the s There appears to be a backlog on articles for review so I guess there n	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero. Any chance you remember what page that's on?	0	0	0	0	0	0
00025465d4725e87	" Congratulations from me as well, use the tools well. · talk "	0	0	0	0	0	0
0002bcb3da6cb337	[redacted] BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to the Matt Shirvington article has been reverted. Plea	0	0	0	0	0	0
00037261f536c51d	Sorry if the word 'nonsense' was offensive to you. Anyway, I'm not inter	0	0	0	0	0	0
00040093b2687caa	alignment on this subject and which are contrary to those of DuLithgow	0	0	0	0	0	0
0005300084f90edc	" Fair use rationale for Image:Wonju.jpg Thanks for uploading Image:Wonju.jpg. I notice the image page specifi Please go to the image description page and edit it to include a fair use If you have uploaded other fair use media, consider checking that you h Unspecified source for Image:Wonju.jpg Thanks for uploading Image:Wonju.jpg. I noticed that the file's descripti As well as adding the source, please add a proper copyright licensing t If you have uploaded other files, consider checking that you have speci	0	0	0	0	0	0
00054a5e18b50dd4	bbq be a man and lets discuss it-maybe over the phone?	0	0	0	0	0	0

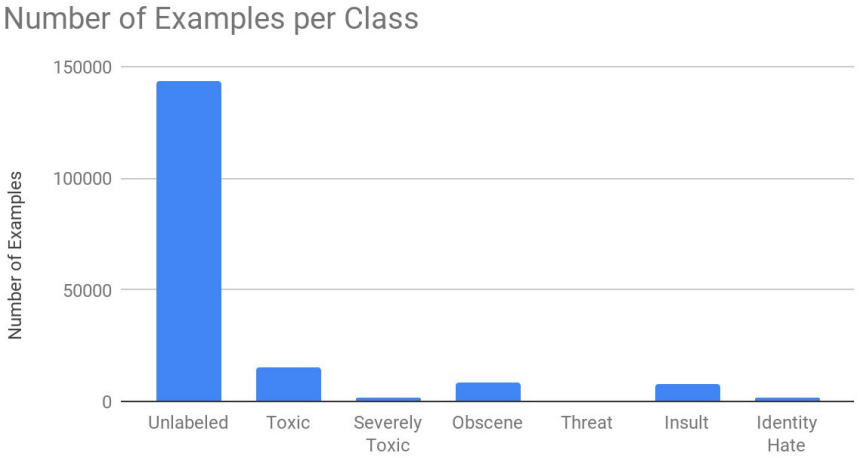
Frameworks

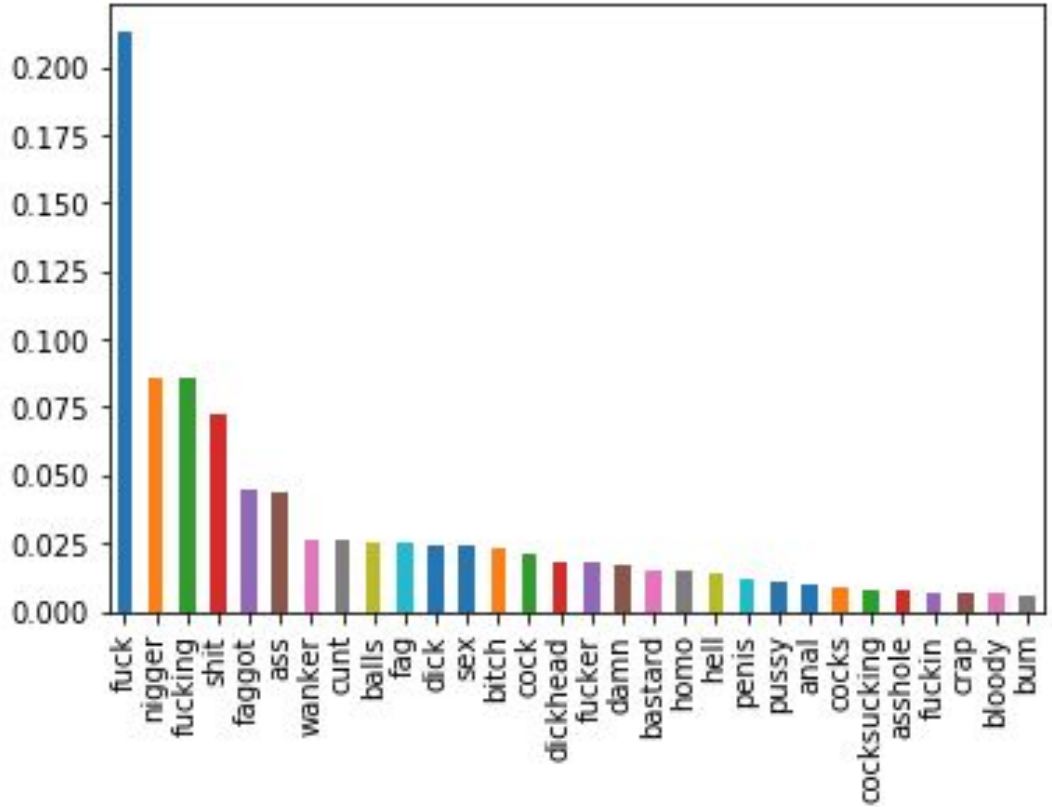
Data Inspection

1. Class Distribution
2. Common Toxic Word Inspection
3. Comment Length Inspection

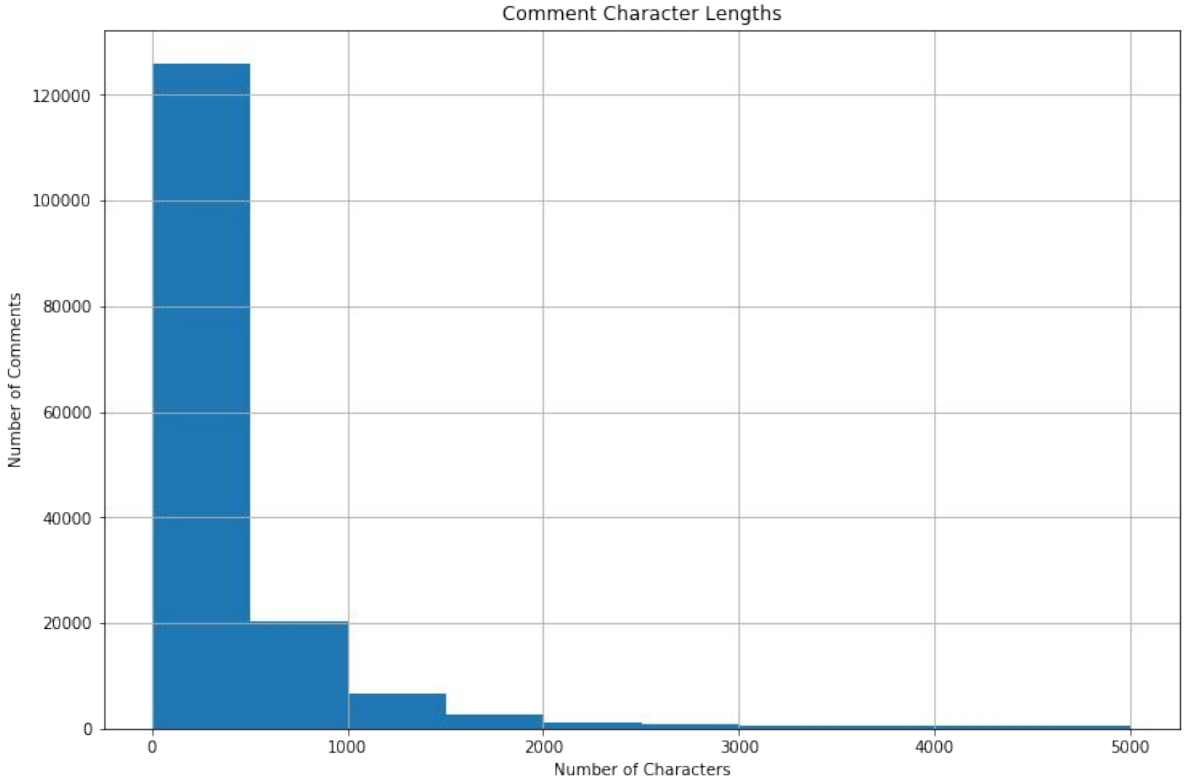
Class Distribution







Common Toxic Words



Comment Character Lengths

Baseline

Random Assignment (based on class frequencies)

	Precision	**Recall**	**F1-Score**	**Support**
Toxic	0.10	0.43	0.16	5038
Severely Toxic	0.01	0.06	0.02	500
Obscene	0.05	0.23	0.08	2810
Threat	0.00	0.02	0.01	152
Insult	0.05	0.23	0.08	2591
Identity Hate	0.01	0.04	0.01	449
Micro Avg	0.07	0.30	0.11	11540
Macro Avg	0.04	0.17	0.06	11540
Weighted Avg	***0.07***	***0.30***	***0.11***	***11540***

Models

1. Naive Bayes Classifier
2. Support Vector Machines
3. Random Forest Classifier
4. Recurrent Neural Network

Naive Bayes Classifier

- “Bag of Words” model makes sense for toxic comment classification
- Precision, Recall, & F1 strong improvements over baseline

	Precision	**Recall**	**F1-Score**	**Support**
Toxic	0.83	0.59	0.69	5042
Severely Toxic	0.31	0.79	0.44	557
Obscene	0.78	0.79	0.79	2761
Threat	0.05	0.78	0.09	163
Insult	0.65	0.68	0.66	2623
Identity Hate	0.19	0.58	0.29	481
Micro Avg	0.53	0.67	0.59	11627
Macro Avg	0.47	0.70	0.49	11627
Weighted Avg	***0.71***	***0.67***	***0.67***	***11627***

Feature Analysis

- Naive Bayes found certain features (unigrams, bigrams, and trigrams) that are most useful to the model

toxic:
- 2123145146
- kundad
- kunstruktive
- kunt
- kupla
- kurang
- yammer
- follarte
- fuckyourself
- crackhead

severe_toxic:
- stomes
- stikin
- caspa
- anastal1111you
- ancest
- ancestryearly
- ancestryerigate
- ada_at
- cartuchos
- homelan

obscene:
- achivements
- achmed
- achsehole
- kcik
- sexmist
- britch
- britbarb
- katzrin
- zigabo
- follarte

threat:
- m45terbate
- ma5terb8
- ma5terbate
- master-bate
- masterb8
- masterbat*
- masterbat3
- teeeccccctoooooniiiiiiccccccc
- hawkinghttp
- zigabo

insult:
- faggots129
- islantic
- snigbrook
- furfag
- fortuijn
- 66185192207
- libtard
- onanizing
- crackhead
- suberbia

identity_hate:
- gomnna
- closerlookonsyria
- nawmean
- goddammed
- clubz
- goains
- nebracka
- negrate
- uos
- zigabo

Support Vector Machines

- Word embeddings to produce embeddings for each sentence
- Leveraged GloVe embeddings
- Leveraging custom embeddings could produce better results with greater resources and greater time

	Precision	Recall	F1-Score	Support
Toxic	0.96	0.06	0.12	6090
Severely Toxic	0.00	0.00	0.00	367
Obscene	0.95	0.09	0.16	3691
Threat	0.00	0.00	0.00	211
Insult	0.67	0.01	0.03	3427
Identity Hate	0.00	0.00	0.00	712
Micro Avg	0.93	0.05	0.10	14498
Macro Avg	0.43	0.03	0.05	14498
Weighted Avg	***0.80***	***0.05***	***0.10***	***14498***

Random Forest Classifier

- Resistant to class imbalance
- Decent results that suffered in the macr average performing poorly in the smaller classes

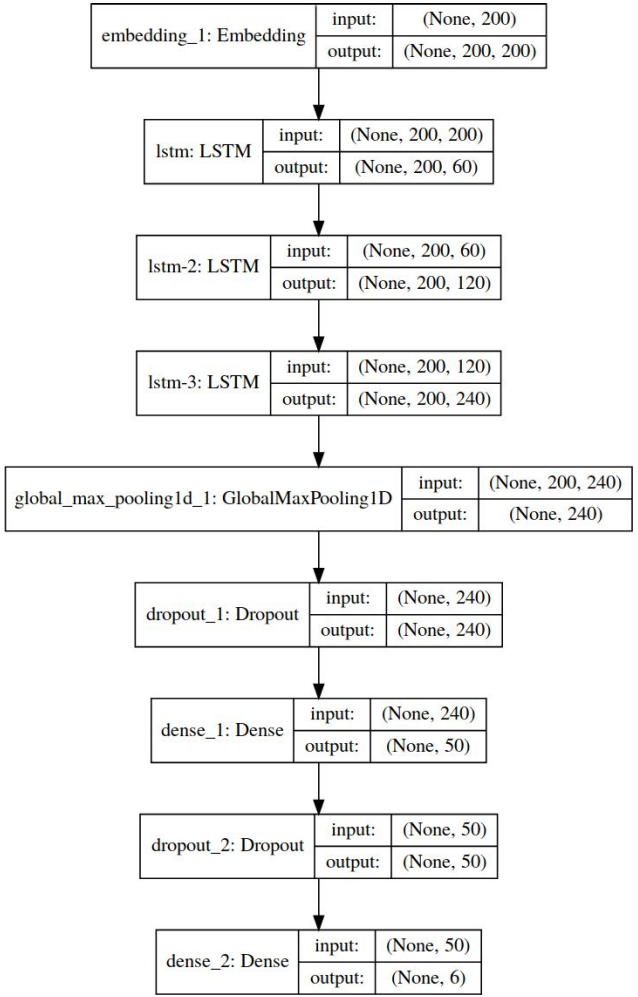
	Precision	Recall	F1-Score	Support
Toxic	0.57	0.76	0.65	6090
Severely Toxic	0.23	0.08	0.12	367
Obscene	0.58	0.68	0.63	3691
Threat	0.33	0.05	0.09	211
Insult	0.56	0.52	0.54	3427
Identity Hate	0.57	0.12	0.20	712
Micro Avg	0.57	0.62	0.59	14498
Macro Avg	0.47	0.37	0.37	14498
Weighted Avg	***0.56***	***0.62***	***0.57***	***14498***

Recurrent Neural Network (RNN)

- LSTMs shown to effectively handle long sequence
- Captures sentence structure

	Precision	Recall	F1-Score	Support
Toxic	0.57	0.85	0.68	6090
Severely Toxic	0.34	0.48	0.40	367
Obscene	0.60	0.80	0.68	3691
Threat	0.00	0.00	0.00	211
Insult	0.52	0.72	0.61	3427
Identity Hate	0.67	0.22	0.34	712
Micro Avg	0.56	0.75	0.64	14498
Macro Avg	0.45	0.51	0.45	14498
Weighted Avg	***0.56***	***0.75***	***0.63***	***14498***

RNN Architecture



Attributions

[1] Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Man´e, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vi´egas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Ga¨el Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122, 2013.

[3] Fran¸cois Chollet et al. Keras. https://keras.io, 2015.

[4] J. D. Hunter. Matplotlib: A 2d graphics environment. Computing In Science & Engineering, 9(3):90–95, 2007.

[5] Wes McKinney. Data structures for statistical computing in python. In St´efan van der Walt and Jarrod Millman, editors, Proceedings of the 9th Python in Science Conference, pages 51 – 56, 2010.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[7] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, pages 1391–1399, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.