Interpreting Neural Network Hate Speech Classifiers

Cindy Wang
Stanford University
Department of Computer Science
cindyw@cs.stanford.edu

Abstract

Deep neural networks have been applied to hate speech detection with apparent success, but they have limited practical applicability without transparency into the predictions they make. In this paper, we perform several experiments to visualize and understand a state-of-the-art neural network classifier for hate speech (Zhang et al., 2018). We adapt techniques from computer vision to visualize sensitive regions of the input stimuli and identify the features learned by individual neurons. We also introduce a method to discover the keywords that are most predictive of hate speech. Our analyses explain the aspects of neural networks that work well and point out areas for further improvement.

1 Introduction

We define hate speech as "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Davidson et al., 2017). This definition importantly does not include all instances of offensive language, reflecting the challenges of automated detection in practice. For instance, in the following examples from Twitter (1) clearly expresses homophobic sentiment, while (2) uses the same term self-referentially:

(1) Being gay aint cool ... yall just be confused and hurt ... **fags** dont make it to Heaven

(2) me showing up in heaven after everyone told me god hates **fags**

As in many other natural language tasks, deep neural networks have become increasingly popular and effective within the realm of hate speech research. However, few attempts have been made to explain the underlying features that contribute to their performance, essentially rendering them black-boxes. Given the significant social, moral, and legal consequences of incorrect predictions, interpretability is critical for deploying and improving these models.

To address this, we contribute three ways of visualizing and understanding neural networks for text classification and conduct a case study on a state-of-the-art model for generalized hate speech. We **1)** perform occlusion tests to investigate regions of model sensitivity in the inputs, **2)** identify maximal activations of network units to visualize learned features, and **3)** identify the unigrams most strongly associated with hate speech. Our analyses explore the bases of the neural network's predictions and discuss common classes of errors that remain to be addressed by future work.

2 Related Work

Hate speech classification. Early approaches employed relatively simple classifiers and relied on manually extracted features (e.g. n-grams, part-of-speech tags, lexicons) to represent data (Dinakar et al., 2011; Nobata et al., 2016). Schmidt and Wiegand (2017)'s survey of hate speech detection describes various types of features used. The classification decisions of such models are interpretable and high-precision: Warner and Hirschberg (2012) found that the trigram "*<DET> jewish <NOUN>*" is the most significant positive feature for anti-semitic hate, while Waseem and Hovy (2016) identified predictive character n-grams using logistic regression coefficients. However, manually extracted feature spaces are limited in both their semantic and syntactic representational ability. Lexical features are insufficient when abusive terms take on various different meanings (Kwok and Wang, 2013; Davidson et al., 2017) or are not present at all in the case of implicit hate speech (Dinakar et al., 2011). Syntactic features such as part-of-speech sequences and typed dependencies fail to fully

Proceedings of the Second Workshop on Abusive Language Online (ALW2), pages 86–92
Brussels, Belgium, October 31, 2018.

capture complex linguistic forms or accurately model context (Waseem and Hovy, 2016; Warner and Hirschberg, 2012). Wiegand et al. (2018) used feature-based classification to build a lexicon of abusive words, which is similar to the interpretability task in this paper of identifying indicative unigram features. While their approach is primarily applicable to explicit abuse, they showed that inducing a generic lexicon is important for cross-domain detection of abusive microposts.

Neural network classifiers. The limitations of feature engineering motivate classification methods that can implicitly discover relevant features. Badjatiya et al. (2017) and Gambäck and Sikdar (2017) were the first to use recurrent neural networks (RNNs) and convolution neural networks (CNNs), respectively, for hate speech detection in tweets. A comprehensive comparative study by Zhang et al. (2018) used a combined CNN and gated recurrent unit (GRU) network to outperform the state-of-the-art on 6 out of 7 publicly available hate speech datasets by 1-13 F1 points. The authors hypothesize that CNN layers capture co-occurring word n-grams, but they do not perform an analysis of the features that their model actually captures. Deep learning classifiers have also been explored for related tasks such as personal attacks and user comment moderation (Wulczyn et al., 2017; Pavlopoulos et al., 2017). Pavlopoulos et al. (2017) propose an RNN model with a self-attention mechanism, which learns a set of weights to determine the words in a sequence that are most important for classification.

Visualizing neural networks. Existing approaches for visualizing RNNs largely involve applying dimensionality reduction techniques such as t-SNE (van der Maaten and Hinton, 2008) to hidden representations. Hermans and Schrauwen (2013) and Karpathy et al. (2015) investigated the functionality of internal RNN structures, visualizing interpretable activations in the context of character-level long short-term memory (LSTM) language models. We are interested in the high-level semantic features identified by network structures and are heavily influenced by the significant body of work focused on visualizing and interpreting CNNs. Zeiler and Fergus (2013) introduced a visualization technique that projects the top activations of CNN layers back into pixel space. They also used partial image occlusion to determine the area of a given image to which the network is most sensitive. Girshick et al. (2013) propose a non-parametric method to visualize learned features of individual neurons for object detection. We adapt these techniques to draw meaningful insights about our problem space.

Label	# Examples	% Examples
Hate	1430	5.8%
Offensive	19190	77.4%
Neither	4163	16.8%

Table 1: Distribution of class labels in the dataset.

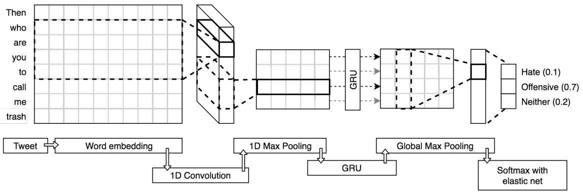


Figure 1: Illustration of the CNN-GRU architecture.

3 Case Study

Dataset. We use the dataset of 24,802 labeled tweets made available by Davidson et al. (2017). The tweets are labeled as one of three classes: hate speech, offensive but not hate speech, or neither offensive nor hate speech. The distribution of labels in the resulting dataset is shown in Table 1. Out of the seven hate speech datasets publicly available at the time of this work,[1] it is the only one that is coded for general hate speech, rather than specific hate target characteristics such as race, gender, or religion.

CNN-GRU model. We utilize the CNN-GRU classifier introduced by Zhang et al. (2018), which achieves the state-of-the-art on most hate speech datasets including Davidson et al. (2017), and contribute a Tensorflow reimplementation for future study. The inputs to the model are tweets which are mapped to sequences of word embeddings. These are then fed through a 1D convolution and max pooling to generate input to a GRU. The output of the GRU is flattened by a global max pooling layer, then finally passed to the softmax output layer, which predicts a probability distribution over the three classes. The model architecture is shown in Figure 1 and described in detail in the original paper.

[1] For details and descriptions of all seven datasets, see Zhang et al. (2018).

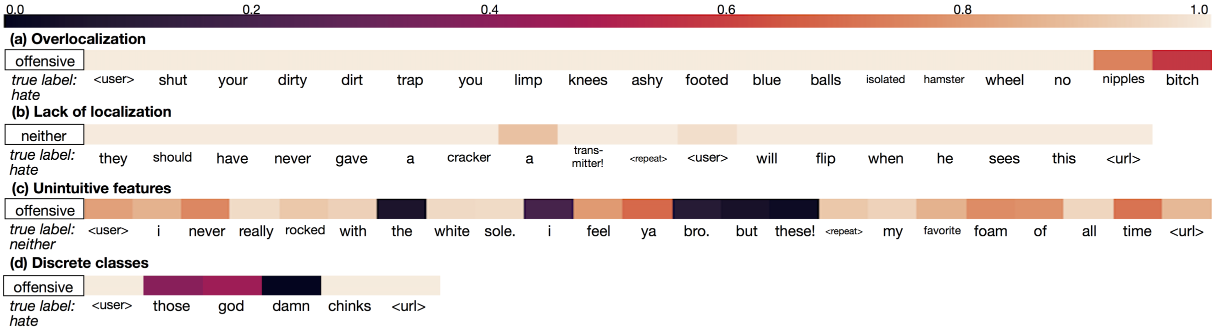


Figure 2: Partial occlusion heatmaps of test examples demonstrating four types of errors made by the CNN-GRU network. Heatmaps are plotted for the predicted class (boxed) while the true class is given below. Darker regions denote portions of the input to which the classifier prediction is most sensitive.

4 Visualization Techniques

4.1 Partial Occlusion

Previously applied to image classification networks, partial occlusion involves iteratively occluding different patches of the input image and monitoring the output of the classifier. We apply a modified version of this technique to our CNN-GRU model by systematically replacing each token of a given input sequence with an `<unk>` token.[2] We then plot a heatmap of the classifier probabilities of the true class (*hate*, *offensive*, or *neither*) over the tokens in the sequence.

The resulting visualizations reveal a few major types of errors made by the CNN-GRU (Figure 2). We observe **overlocalization** in many long sequences, particularly those misclassified as offensive. This occurs when the classifier decision is sensitive to only a single unigram or bigram rather than the entire context, as in Figure 2(a). The network loses sequence information during convolution and shows decreased sensitivity to the longer context as a result.

Lack of localization is the opposite problem in which the model is not sensitive to any region of the input, shown in Figure 2(b). It occurs mostly in longer and hate class examples. A possible explanation for this type of error is that convolving sequential data diffuses the signal of important tokens and n-grams.

For correctly classified examples, the sensitive regions intuitively correspond to features like n-grams, part-of-speech templates, and word dependencies. However, incorrectly classified examples also demonstrate sensitivity to **unintuitive features** that are not helpful for classification. For instance, Figure 2(c) shows a sensitive region that crosses a sentence boundary.

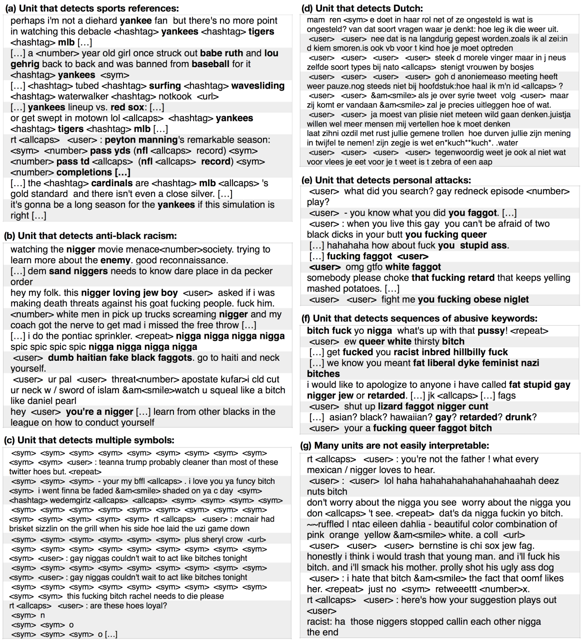
Finally, we see a high rate of errors due to the **discretization of the *hate* and *offensive* classes.** While hate speech is largely contained within offensive language, the sensitive regions for the two classes are disparate. In Figure 2(d), the network's prediction that the example is offensive is highly sensitive to the sequence *"those god damn"*, but not the racial slur *"chinks,"* which is both hateful and offensive.

Some of the errors we identify, such as lack of localization and unintuitive features, can be addressed by modifying the model architecture. We can change the way long sequences are convolved, or restrict convolutions within phrase boundaries. More difficult to address are the errors in which our network is sensitive to the correct regions (or a reasonable subset thereof) but makes incorrect predictions. These issues stem from the nature of the data itself, such as the complex linguistic similarity between hate and offensive language. Moreover, many misclassified examples contain implicit characteristics such as sarcasm or irony, which limits the robustness of features learned solely from input text.

4.2 Maximum Activations

The technique of retrieving inputs that maximally activate individual neurons has been used for image networks (Zeiler and Fergus, 2013; Girshick et al., 2013), and we adapt it to the CNN-GRU in order to understand what features of the input stimuli it learns to detect. For each of the units in the final global max pooling layer of the CNN-

[2] `<unk>` indicates an out-of-vocabulary word. The word embedding for `<unk>` is random, whereas in-vocabulary word embeddings encode meaning via unsupervised pretraining.



(a) Unit that detects sports references:

perhaps i'm not a diehard **yankee** fan but there's no more point in watching this debacle <hashtag> **yankees** <hashtag> **tigers** <hashtag> **mlb** […]

[…] a <number> year old girl once struck out **babe ruth** and **lou gehrig** back to back and was banned from **baseball** for it <hashtag> **yankees** <sym>

[…] <hashtag> tubed <hashtag> **surfing** <hashtag> **wavesliding** <hashtag> waterwalker <hashtag> notkook <url>

[…] **yankees** lineup vs. **red sox**: […]

or get swept in motown lol <allcaps> <hashtag> **yankees** <hashtag> **tigers** <hashtag> **mlb** […]

rt <allcaps> <user> : **peyton manning**'s remarkable season: <sym> <number> **pass yds** (**nfl** <allcaps> record) <sym> <number> **pass td** <allcaps> (**nfl** <allcaps> **record**) <sym> <number> **completions […]**

[…] the <hashtag> **cardinals** are <hashtag> **mlb** <allcaps> 's gold standard and there isn't even a close silver. […]

it's gonna be a long season for the **yankees** if this simulation is right […]

(b) Unit that detects anti-black racism:

watching the **nigger** movie menace<number>society. trying to learn more about the **enemy**. good reconnaissance.

[…] dem **sand niggers** needs to know dare place in da pecker order

hey my folk. this **nigger loving jew boy** <user> asked if i was making death threats against his goat fucking people. fuck him.

<number> white men in pick up trucks screaming **nigger** and my coach got the nerve to get mad i missed the free throw […]

[…] i do the pontiac sprinkler. <repeat> **nigga nigga nigga nigga** spic spic spic spic **nigga nigga nigga nigga**

<user> **dumb haitian fake black faggots**. go to haiti and neck yourself.

<user> ur pal <user> threat<number> apostate kufar>i cld cut ur neck w / sword of islam &am<smile>watch u squeal like a bitch like daniel pearl

hey <user> **you're a nigger** […] learn from other blacks in the league on how to conduct yourself

(c) Unit that detects multiple symbols:

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : teanna trump probably cleaner than most of these twitter hoes but. <repeat>

<sym> <sym> <sym> - your my bffl <allcaps> . i love you ya funcy bitch <sym> i went finna be faded &am<smile> shaded on ya c day <sym> <hashtag> wedemgirlz <allcaps> <allcaps> <sym> <sym> <sym> <sym>

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> rt <allcaps> <user> : mcnair had brisket sizzlin on the grill when his side hoe laid the uzi game down

<sym> <sym> <sym> <sym> <sym> <sym> plus sheryl crow <url>

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : gay niggas couldn't wait to act like bitches tonight

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : gay niggas couldn't wait to act like bitches tonight

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> this fucking bitch rachel needs to die please

rt <allcaps> <user> : are these hoes loyal?
<sym> n
<sym> <sym> o
<sym> <sym> <sym> o […]

(d) Unit that detects Dutch:

mam ren <sym> e doet in haar rol net of ze ongesteld is wat is ongesteld? van dat soort vragen waar je denkt: hoe leg ik die weer uit.

<user> <user> nee dat is na langdurig gepest worden.zoals ik al zei:in d kiem smoren.is ook vb voor t kind hoe je moet optreden

<user> <user> <user> <user> steek d morele vinger maar in j neus zelfde soort types bij nato <allcaps> stenigt vrouwen by bosjes

<user> <user> <user> <user> goh d anoniemeaso meeting heeft weer pauze.nog steeds niet bij hoofdstuk:hoe haal ik m'n id <allcaps> ?

<user> <user> &am<smile> als je over syrie tweet volg <user> maar zij komt er vandaan &am<smile> zal je precies uitleggen hoe of wat.

<user> <user> ja moest van plisie niet meteen wild gaan denken.juistja willen wel meer mensen mij vertellen hoe k moet denken

laat zihni ozdil met rust jullie gemene trollen hoe durven jullie zijn mening in twijfel te nemen! zijn zegje is wet en*kuch**kuch*. .water

<user> <user> <user> <user> tegenwoordig weet je ook al niet wat voor vlees je eet voor je t weet is t zebra of een aap

(e) Unit that detects personal attacks:

<user> what did you search? gay redneck episode <number> play?

<user> - you know what you did **you faggot**. […]

<user> : when you live this gay you can't be afraid of two black dicks in your butt **you fucking queer**

[…] hahahaha how about fuck **you stupid ass**.

[…] **fucking faggot <user>**

<user> omg gtfo **white faggot**

somebody please choke **that fucking retard** that keeps yelling mashed potatoes. […]

<user> <user> fight me **you fucking obese niglet**

(f) Unit that detects sequences of abusive keywords:

bitch fuck yo **nigga** what's up with that **pussy**! <repeat>

<user> ew **queer white** thirsty **bitch**

[…] get **fucked** you **racist inbred hillbilly fuck**

[…] we know you meant **fat liberal dyke feminist nazi bitches**

i would like to apologize to anyone i have called **fat stupid gay nigger jew** or **retarded**. […] jk <allcaps> […] fags

<user> shut up **lizard faggot nigger cunt**

[…] asian? black? hawaiian? **gay**? **retarded**? **drunk**?

<user> your a **fucking queer faggot bitch**

(g) Many units are not easily interpretable:

rt <allcaps> <user> : you're not the father ! what every mexican / nigger loves to hear.

<user> : <user> lol haha hahahahahahahahahaahah deez nuts bitch

don't worry about the nigga you see worry about the nigga you don <allcaps> 't see. <repeat> dat's da nigga fuckin yo bitch.

~~ruffled l ntac eileen dahlia - beautiful color combination of pink orange yellow &am<smile> white. a coll <url>

<user> <user> <user> bernstine is chi sox jew fag.

honestly i think i would trash that young man. and i'll fuck his bitch. and i'll smack his mother. prolly shot his ugly ass dog

<user> : i hate that bitch &am<smile> the fact that oomf likes her. <repeat> just no <sym> retweeettt <number>x.

rt <allcaps> <user> : here's how your suggestion plays out
<user>
racist: ha those niggers stopped callin each other nigga
the end

Figure 3: Examples of interpretable units from the global max pool layer of the CNN-GRU. The inputs with the top eight activations for each neuron are shown, with relevant tokens bolded. In the interest of space, some examples here are abridged and the full version can be found in Appendix Figure 4.

GRU, we compute the unit's activations on the entire dataset and retrieve the top-scoring inputs.

Figure 3 displays the maximally activating examples for seven of 100 units in the global max pool. We verify that the model does indeed learn relevant features for hate speech classification, some of which are traditional natural language features such as the part-of-speech bigram *"you <NOUN PHRASE>."* Others, like a unit that fires on sports references and a unit that detects Dutch-language tweets (the result of querying for hate keywords *yankee* and *hoe,* respectively), reflect assumptions in data collection. Some units capture features that reflect domain-specific phenomena, such as repeated symbols or sequences of multiple abusive keywords.

Many units are too general or not interpretable at all. For instance, several units detect the hate term *bitch,* but none of them clearly capture the distinction between when it is used in sexist and colloquial contexts. Conversely, examples containing rarer and more ambiguous slurs like *cracker* do not appear as top inputs for any unit. Overall, the CNN-GRU discovers some interpretable lexical and syntactical features, but its final layer activations overrepresent common unigrams and fail to detect semantics at a more fine-grained level via surrounding context.

Category	Terms
Hatebase words	faggot, nigger, fag, coon, teabagger, cripple, spook, muzzie, mook, jiggaboo, mutt, redskin, dink
Hatebase plural	faggots, niggers, fags, crackers, coons, rednecks, hos, queers, coloreds, wetbacks, muzzies, wiggers, darkies
Pejorative nominalization	blacks, jews, commie, lightskins, negroes
Hate-related or offensive	racist, fugly, hag, traitor, chucks, goon, asss, blacc, eff, homophobic, racists, nogs, muhammed, fatherless, slurs
Hateful context words	arrested, yelled, smoked, joints, stoned, frat, celibate, catfished, wedlock, sliced, kappa, trappin, birthed, allegiance, menace, commander, stamp, cyber
Hashtags	*(see Table 4)*
Dialect variations	des, boutta, denna, waddup, boof, ont, bestf, playen, sav, erbody, prolli, deze, bougie
Pop culture	gram, tweakin, dej, uchiha, mewtwo, bios, fenkell, mikes, beavis, aeropostale
Other	en, waffle, moe, saltine, squid, pacer, sharpie, skyler, sockfetish, johns, lactose, ov, tater

Table 2: List of keywords discovered via synthetic test examples. Terms within each category appear in order of frequency in the corpus. Descriptions for hashtags are from blogs such as www.socialseer.com and cross-referenced on Twitter.

4.3 Synthetic Test Examples

Lastly, we propose a general technique to identify the the most indicative unigram features for a deep model using synthetic test examples and apply it to the CNN-GRU.

For each word in our corpus, we construct a sentence of the form *"they call you* `<word>`*"* and feed it as input to the CNN-GRU network. We choose this structure to grammatically accommodate both nouns and adjectives, and because it is semantically neutral compared to similar formulations such as *"you are* `<word>`*."* We then retrieve the words whose test sentences are classified as hate speech. After filtering out words that do not appear in two or more distinct tweets (retweets are indistinct) and words containing non-alphabetical characters,[3] this method returns 101 terms. We manually group the terms into nine categories and

[3] We inspect the output to confirm that this filtering eliminates only nonsense tokens and not intentional misspellings.

summarize them in Table 2.

As a quantitative heuristic for the quality of the discovered terms, we evaluate our method's recall on the hate speech lexicon Hatebase.[4] We measure the recall of Hatebase words, plural forms of Hatebase words, and tweets containing Hatebase terms and compare against a random baseline (Table 3). The recall of our method is approximately an order of magnitude better than random across all categories. Recall of plural forms is better than that of base forms, which may reflect the training data's definition of hate speech as language that targets a group. Notably, recall of Hatebase tweets[5] is lower than recall of individual terms regardless of form, meaning that the Hatebase terms discovered using our template method are not the ones that occur most commonly in the corpus. This is reasonable as several of the most common Hatebase terms such as *bitch* and *nigga* are ones that tend to be used colloquially rather than as slurs.

Of the non-Hatebase terms that our method discovers, four are pejorative nominalizations. These are neutral adjectives that take on pejorative meaning when used as nouns, such as *blacks* and *jews* (Palmer et al., 2017). We also find six domain-specific hate terms in the form of hashtags, which tend to be non-word acronyms and primarily used by densely connected, politically conservative Twitter users (see Table 4). The results also include dialect-specific terms and slang spellings, such as *des* and *boutta*, which mean *these* and *about to* respectively. While these terms co-occur frequently with hate speech keywords in our corpus, they are semantically neutral, suggesting that our model exhibits bias towards certain written vernaculars. While these terms co-occur frequently with hate speech keywords in our corpus, they are semantically neutral, suggesting that our model exhibits bias towards certain written vernaculars.

	Recall	Recall %	Recall %
	(ours)	(ours)	(random)
Hb terms	13/182	7.14	1.05
Hb plurals	13/91	14.29	1.06
Both	26/273	9.52	1.04
Hb tweets	1453/24234	6.00	1.21

Table 3: Comparison between our method and a random baseline on recall of Hatebase lexicon terms and tweets containing Hatebase terms. The random baseline is averaged over 10,000 trials.

Hashtag	Meaning
#pjnet	Patriot Journalist Network
#lnyhbt	Let Not Your Heart Be Troubled, Sean Hannity's hashtag
#tgdn	Twitter Gulag Defense Network
#httr	Hail to the Redskins
#yeswedid	Reference to President Obama's motto
#acab	All Coppers Are Bastards

Table 4: Descriptions of discovered hashtag keywords, given by blogs such as www.socialseer.com and cross-referenced on Twitter.

5 Conclusion

We presented a variety of methods to understand the prediction behavior of a neural network text classifier and applied them to hate speech. First, we used partial occlusion of the inputs to visualize the sensitivity of the network. This revealed that the architecture loses representational capacity on long inputs and suffers from lack of class separability. We then analyzed the semantic meaning of individual neurons, some of which capture intuitively good features for our domain, though many still appear to be random or uninterpretable. Finally, we presented a way to discover the most indicative hate keywords for our model. Not all discovered terms are inherently hateful, revealing peculiarities and biases of our problem space.

Overall, our experiments give us better insight into the implicit features learned by neural networks. We lay the groundwork for future efforts towards better modeling and data collection, including active removal of linguistic discrimination.

Acknowledgments

Thanks firstly to Christopher Potts for many useful discussions that formed the foundation for this work in this paper. Thanks also to the three anonymous reviewers for their insightful feedback and constructive suggestions. This material is based in part upon work supported by the NSF under Grant No. BCS-1456077.

[4] Hatebase (https://www.hatebase.org) is an online, crowd-sourced hate speech lexicon. Davidson et al. (2017) use Hatebase queries to bootstrap the dataset we use in this paper.

[5] The number of tweets containing one of the 26 discovered terms divided by the number of tweets containing any Hatebase term.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *WWW 2017 Companion*, pages 759–760. International World Wide Web Conference Committee.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pages 512–515.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Fifth International AAAI Conference on Weblogs and Social Media Workshop on the Social Mobile Web*, pages 11–17.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Michiel Hermans and Benjamin Schrauwen. 2013. Training and analysing deep recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 190–198.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. *CoRR*, abs/1506.02078.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1621–1622. Association for the Advancement of Artificial Intelligence.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153.

Alexis Palmer, Melissa Robinson, and Kristy Phillips. 2017. Illegal is not a noun: Linguistic form for detection of pejorative nominalizations. In *Proceedings of the First Workshop on Abusive Language Online*, pages 91–100. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2012 Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 88–93. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1046–1056. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference 2018*, pages 745–760.

A Appendix

		Predicted labels		
		Hate	Offensive	Neither
True labels	Hate	0.31	0.58	0.11
	Offensive	0.04	0.94	0.03
	Neither	0.02	0.13	0.85

Table 5: A summary of performance for the CNN-GRU classifier on a held-out test set.

(a) Unit that detects sports references:

perhaps i'm not a diehard **yankee** fan but there's no more point in watching this debacle <hashtag> **yankees** <hashtag> **tigers** <hashtag> **mlb** <allcaps> <allcaps> so what else is on? lol <allcaps>

rt <allcaps> <user> : a <number> year old girl once struck out **babe ruth** and **lou gehrig** back to back and was banned from **baseball** for it <hashtag> **yankees** <sym>

<hashtag> mt <hashtag> commission <hashtag> gouache <hashtag> tiki <hashtag> wahine <hashtag> monkey <hashtag> tubed <hashtag> **surfing** <hashtag> **wavesliding** <hashtag> waterwalker <hashtag> notkook <url>

rt <allcaps> <user> : <hashtag> **yankees** lineup vs. **red sox**: […]

or get swept in motown lol <allcaps> <hashtag> **yankees** <hashtag> **tigers** <hashtag> **mlb** <allcaps> <allcaps>

rt <allcaps> <user> : **peyton manning**'s remarkable season:
<sym> <number> **pass yds** (**nfl** <allcaps> record)
<sym> <number> **pass td** <allcaps> (**nfl** <allcaps> **record**)
<sym> <number> **completions**
<sym> likely ho <sym>

rt <allcaps> <user> : the <hashtag> **cardinals** are <hashtag> mlb <allcaps> 's gold standard and there isn't even a close silver. <hashtag> **dodgers** <hashtag> **giants** <hashtag> **tigers** <hashtag> **yankees** <url> <sym>

it's gonna be a long season for the **yankees** if this simulation is right <hashtag> **bluejays** <hashtag> **rays** <hashtag> **yankees** <hashtag> **orioles** <sym> <url>

(b) Unit that detects anti-black racism:

watching the **nigger** movie menace<number>society. trying to learn more about the **enemy**. good reconnaissance.

<user> <user>
i likes dat name. dem **sand niggers** needs to know dare place in da pecker order

hey my folk. this **nigger loving jew boy** <user> asked if i was making death threats against his goat fucking people. fuck him.

<number> white men in pick up trucks screaming **nigger** and my coach got the nerve to get mad i missed the free throw <sym> <sym> <sym>

<hashtag> tweetlikepontiacholmes <sym>
i do the pontiac sprinkler. <repeat>
nigga nigga nigga nigga spic spic spic spic **nigga nigga nigga nigga**

<user> **dumb haitian fake black faggots**. go to haiti and neck yourself.

<user> ur pal <user> threat<number> apostate kufar>i cld cut ur neck w / sword of islam &am<smile>watch u squeal like a bitch like daniel pearl

hey <user> **you're a nigger** <hashtag> racismisaliveandwellbro   learn from other blacks in the league on how to conduct yourself

(c) Unit that detects multiple symbols:

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : teanna trump probably cleaner than most of these twitter hoes but. <repeat>

<sym> <sym> <sym> - your my bffl <allcaps> . i love you ya funcy bitch <sym> i went finna be faded &am<smile> shaded on ya c day <sym> <hashtag> wedemgirlz <allcaps> <allcaps> <sym> <sym> <sym> <sym>

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> rt <allcaps>   <user> : mcnair had brisket sizzlin on the grill when his side hoe laid the uzi game down

<sym> <sym> <sym> <sym> <sym> <sym> plus sheryl crow <url>

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : gay niggas couldn't wait to act like bitches tonight

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <user> : gay niggas couldn't wait to act like bitches tonight

<sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> <sym> this fucking bitch rachel needs to die please

rt <allcaps>   <user> : are these hoes loyal?
<sym> n
<sym> <sym> o
<sym> <sym> <sym> o
<sym> <sym> <sym> <sym> o […]

(d) Unit that detects Dutch:

mam ren <sym> e doet in haar rol net of ze ongesteld is wat is ongesteld? van dat soort vragen waar je denkt: hoe leg ik die weer uit.

<user> <user> nee dat is na langdurig gepest worden.zoals ik al zei:in d kiem smoren.is ook vb voor t kind hoe je moet optreden

<user> <user> <user> <user> steek d morele vinger maar in j neus zelfde soort types bij nato <allcaps> stenigt vrouwen by bosjes

<user> <user> <user> <user> goh d anoniemeaso meeting heeft weer pauze.nog steeds niet bij hoofdstuk:hoe haal ik m'n id <allcaps> ?

<user> <user> &am<smile> als je over syrie tweet volg <user> maar zij komt er vandaan &am<smile> zal je precies uitleggen hoe of wat.

<user> <user> ja moest van plisie niet meteen wild gaan denken.juistja willen wel meer mensen mij vertellen hoe k moet denken

laat zihni ozdil met rust jullie gemene trollen hoe durven jullie zijn mening in twijfel te nemen! zijn zegje is wet en*kuch**kuch*. .water

<user> <user> <user> tegenwoordig weet je ook al niet wat voor vlees je eet voor je t weet is t zebra of een aap

(e) Unit that detects personal attacks:

<user> what did you search? gay redneck episode <number> play?

<user> - you know what you did **you faggot**. there are literally thousands of people that have you on their radar homo

<user> : when you live this gay you can't be afraid of two black dicks in your butt **you fucking queer**

lol so since i'm from the south i automatically sound like a fucking redneck. hahahaha how about fuck **you stupid ass**.

i bet you think you a no-chill savage huh. **fucking faggot <user>**

<user> omg gtfo **white faggot**

somebody please choke that fucking retard that keeps yelling mashed potatoes. what a tool! <hashtag> pgac <allcaps> hampionship

<user> <user> fight me **you fucking obese niglet**

(f) Unit that detects sequences of abusive keywords:

bitch fuck yo **nigga** what's up with that **pussy**! <repeat>

<user> ew **queer white** thirsty **bitch**

rt <allcaps> <user> : <user> <user> <user> get **fucked** you **racist inbred hillbilly fuck**

<user> done that spell check we know you meant **fat liberal dyke feminist nazi bitches**

i would like to apologize to anyone i have called **fat stupid gay nigger jew** or **retarded**. i am truly sorry. <repeat>
jk <allcaps>
just trolling
fags

<user> shut up **lizard faggot nigger cunt**

<hashtag> somethingig <allcaps> etalot are you. <repeat> asian? black? hawaiian? **gay**? **retarded**? **drunk**?

<user> your a **fucking queer faggot bitch**

(g) Many units are not easily interpretable:

rt <allcaps> <user> : you're not the father ! what every mexican / nigger loves to hear.

<user> : <user> lol haha hahahahahahahahahaahah deez nuts bitch

don't worry about the nigga you see worry about the nigga you don <allcaps> 't see. <repeat> dat's da nigga fuckin yo bitch.

∼∼ruffled | ntac eileen dahlia - beautiful color combination of pink orange yellow &am<smile> white. a coll <url>

<user> <user> <user> bernstine is chi sox jew fag.

honestly i think i would trash that young man. and i'll fuck his bitch. and i'll smack his mother. prolly shot his ugly ass dog

<user> : i hate that bitch &am<smile> the fact that oomf likes her. <repeat> just no <sym> retweeettt <number>x.

rt <allcaps> <user> : here's how your suggestion plays out <user>
racist: ha those niggers stopped callin each other nigga
the end

Figure 4: Version of Figure 3 with full text of each tweet. Examples of interpretable units from the global max pool layer of the CNN-GRU. The inputs with the top eight activations for each neuron are shown, with relevant tokens bolded.