Incivility Detection in Online Comments

Farig Sadeque
School of Information
University of Arizona
Tucson, AZ 85721
farig@email.arizona.edu

Stephen Rains
Dept. of Communication
University of Arizona
Tucson, AZ 85721
srains@email.arizona.edu

Yotam Shmargad
School of Govt. and Public Policy
University of Arizona
Tucson, AZ 85721
yotam@email.arizona.edu

Kate Kenski
Dept. of Communication
University of Arizona
Tucson, AZ 85721
kkenski@email.arizona.edu

Kevin Coe
Dept. of Communication
University of Utah
Salt Lake City, UT 84112
kevin.coe@utah.edu

Steven Bethard
School of Information
University of Arizona
Tucson, AZ 85721
bethard@email.arizona.edu

Abstract

Incivility in public discourse has been a major concern in recent times as it can affect the quality and tenacity of the discourse negatively. In this paper, we present neural models that can learn to detect name-calling and vulgarity from a newspaper comment section. We show that in contrast to prior work on detecting toxic language, fine-grained incivilities like name-calling cannot be accurately detected by simple models like logistic regression. We apply the models trained on the newspaper comments data to detect uncivil comments in a Russian troll dataset, and find that despite the change of domain, the model makes accurate predictions.

1 Introduction

Online harassment, colloquially known as cyberbullying or cyber harassment, has been rampant since the introduction of the Internet to the general population. It has been a major cause of concern since the mid- and late-90's, and is a thoroughly researched topic in the fields of social science, behavioral science, network science and computer security. Cyberbullying is a form of harassment that is carried out using electronic modes of communication like computer, phone, and in almost all the cases in recent years, the Internet. Cyberbullying is defined as a "willful and repeated harm inflicted through the medium of electronic text" by Patchin and Hinduja (2006)- but this phenomenon goes far beyond the scope of just electronic text. A more comprehensive definition of cyberbullying can be found in one of their later works, where they defined cyberbullying as "a form of harassment using electronic mode of communication" (Hinduja and Patchin, 2008). Fauman (2008) described cyberbullying as "bullying through the use of technology such as the Internet and cellular phones".

The spectrum of online harassment is vast; hence, we focus on one segment of this phenomenon: online incivility. Incivility has been rampant in American society for quite some time. Incivility is described as *features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics* (Coe et al., 2014). While it is often said that incivility is "very much in the eye of the beholder" and what is civil to someone may be uncivil to another (Kenski et al., 2017), some are universal nevertheless. One study has suggested that 69% of Americans believe that incivility in public discourse has become a rampant problem, and only 6% do not identify it as a problem (Shandwick, 2018). The average number of incivility encounters per week has also risen drastically in both the physical world and cyberspace. Social media encounters are especially alarming: a person who encountered any form of incivility anywhere, had on average 5.4 uncivil encounters per week in online social media platforms in 2018, which is almost double the amount from late 2016.

In this paper, we present machine learning models that can identify two prominent forms of incivility, name-calling and vulgarity, based on user-generated contents from public discourse platforms. We focused trained recurrent neural network models on an annotated newspaper comment section and showed that our model outperforms several baselines, including a state-of-the-art model based

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 283–291
Minneapolis, June 6–7, 2019.

on pre-trained contextual embeddings. We applied our newspaper-comments-trained model to a datsaets of Russian troll tweets to observe how the model generalizes from one platform to another.

2 Related Works

Kenski et al. (2017) divided incivility into several different forms, including name-calling, vulgarity, lying accusation, pejorative, and aspersion. They took comments posted by regulars in a newspaper website, and annotated these for the various forms of incivility. Their research focused mostly on the demographics and other individual attributes of readers of these comments and how they perceived incivility in these comments.

Rains et al. (2017) focused more on the perpetrators of incivility rather than the readers. They researched a handful of news articles published in the Arizona Daily Star newspaper website and the comments posted about these articles, then manually annotated these comments and their posters for their incivility and political orientation. The authors found that conservatives were significantly less likely to be uncivil in these public discussions compared to liberals, and the likelihood of liberals being uncivil increased with the presence of conservatives in the same discussion. Liberals were also found to be more repercussive compared to the conservatives.

Recent work has focused on particular forms of incivility, as described in the following sections.

2.1 Generic incivility

Reynolds et al. (2011) developed machine learning models that can detect cyberbullying by identifying curse and insult words in social media posts. They have collected a small set of posts from a website named *formspring.me* and used various non-sequential learning algorithms on this dataset to build a binary classifier for cyberbullying detection.

2.2 Vulgarity

Cachola et al. (2018) used a vulgarity score for better sentiment prediction from a collection of 6800 tweets. They found that vulgarity interacts with key demographic variables like gender, age, religiosity, etc. Other research has also identified demographic keys closely associated with vulgarity: Wang et al. (2014) presented a quantitative analysis on the frequency of curse word usage in Twitter and their variation with certain demographics, and Gauthier et al. (2015) analyzed the usage of swear words based on Tweeter users' age and gender. As none of these papers present any machine learning model that can be used for vulgarity detection, Holgate et al. (2018) claim their work to be the first in vulgarity prediction. They classified functionality of vulgarity in five different cohorts: aggression, emotion expression, emphasis, auxiliary and signalling group identity; and used binary logistic regression classifiers to identify vulgar texts. They also showed the correlation among demographic variables and vulgarity and found that age, faith, and political ideology have significant correlation with vulgarity usage.

2.3 Racism/sexism

Waseem and Hovy (2016) has presented machine learning models that can be used to detect racism and sexism in social media. They have collected and annotated a set of almost 17000 tweets, and used them to build character based n-gram models for offensive tweet detection. They have provided an extensive list of criteria that identify a tweet as racially and sexually offensive, and showed that demographic information does not add much performance to a character-level model.

2.4 Personal attacks

Wulczyn et al. (2017) introduced a methodology to generate annotations for personal attacks. They have used crowdsourcing to identify a set of Wikipedia comments, and used a machine learning model to imitate this annotation on a much larger scale. Agrawal and Awekar (2018) have developed deep neural models that can detect cyberbullying (Reynolds et al., 2011), racism/sexism (Waseem and Hovy, 2016), and personal attacks (Wulczyn et al., 2017) in multiple social media platforms. They claim that theirs is the first work to systematically analyze cyberbullying in social media towards building deep prediction models. They have shown that hand-crafted features using lexicons is not a good idea as abusive word vocabularies vary a lot from one social media platform to another, and swear words are not always considered to be uncivil in social media.

2.5 Name-calling

Habernal et al. (2018) analyzed ad hominem attacks in *Change My View*, a "good faith" argumentation platform that is hosted on Reddit.

They identified posts that Reddit moderators had marked as violating the forum's rules against ad hominem atacks. To identify such posts, they used stacked bidirectional Long-Short Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs), and achieved 78% and 81% accuracy, respectively. One of their most interesting findings was that in 48.6% of the cases, ad hominem attacks are in the last comment of the thread, which shows that personal attacks and name-callings can affect user participation in public discourses.

Works that closely resemble what we are trying to do have one major issue with the datasets that have been used- they are often annotated by mechanical turks (Wulczyn et al., 2017; Reynolds et al., 2011). Incivility is based on the perception of the person in the receiving end, and this perception varies wildly from person to person. Using turkers that we know almost nothing about is not ideal- as difference in perception may introduce unintended bias in the dataset. Hence, we need a dataset that is annotated by experts who have extensive knowledge on incivility detection. Coe et al. (2014) presents one such dataset, and we plan to use this for our incivility detection task (more on this in Section 4).

3 Incivility Classification and Definitions

For our work, we will use the incivility classification presented by Coe et al. (2014): name-calling, vulgarity, aspersion, lying accusations and pejorative for speech. We focus on the two most prevalent forms of these in Coe et al. (2014)'s data: name-calling and vulgarity.

name-calling Ad hominem attacks. Although ad hominem attacks are often used to derail a conversation by using derogatory terms towards another person, the authors have included every instances of derogatory remarks, irrespective of target and intention. For example, *At least the morons in the state capital no longer have control of this process!* is identified as a name-calling comment as it has the word *moron* in it (Kenski et al., 2017).

vulgarity Contents that include any sort of curse words, including minor ones such as *damn* (Kenski et al., 2017). For example, *I hope the voters will kick that politician out on his pompous ass next election.* is marked as vulgar, as it contains the word *ass* in it.

4 Data

Coe et al. (2014) graciously shared with us the data that they collected from the comment section of the *Arizona Daily Star* newspaper. They collected articles and comments between 17 October 2011 and 6 November 2011 from eight news sections: Business, Entertainment, Lifestyles, Local News, Nation and World, Opinion, Sports, and State News. All their data was downloaded and saved manually by one research assistant one day after the articles were posted to provide enough time for the article to garner comments, yet not long enough for the article to be deleted. At the end of the data collection period, a total of 706 articles and 6535 comments were collected, out of which they coded 6444 for further analysis.

They used three teams of 3-5 research assistants to code articles and comments for incivility. The teams had extensive training on the coding procedures (Coe et al., 2014). The coding process took approximately six weeks, and chance-corrected intercoder reliability was established prior to the coding, which ranged between 0.61 to 1.0 Krippendorff's alpha score for different codes. In addition to coding the incivilities present in the comments, they also coded a variety of other metadata, e.g., the author's name, reactions received for other readers (thumbs up or thumbs down), word counts, etc. All the results of the coding procedure were saved in a metadata file created using Microsoft Excel. Comments were saved in separate PDF files named based on the news sections, articles and dates.

5 Challenges in Identifying incivilities from User Contents

As we have mentioned before that incivility is in the eye of the beholder, it is sometimes challenging to identify what can be unequivocally considered as uncivil interaction. Informed by the Coe et al. (2014) data, the following sections discuss some of these challenges.

5.1 Frequency

Although researchers have identified incivilities being rampant in public discourse (Shandwick, 2018), it is still minuscule compared to regular civil discourses in any social platform. As most of our identification and prediction techniques are data-driven, it is difficult to create a model that can identify incivilities from this small number of examples.

5.2 Linguistic Variations and Creativity

Oftentimes people refrain from using an exact version of an uncivil phrase and use an abbreviation or spelling variation of that phrase instead. For example, in *All BS, just like the politicians – the same crap*, the term BS is clearly an abbreviation of the word *bullshit*. However, there are also instances in the data where BS is used to abbreviate a person's name, which clearly is not an example of uncivil comment. Also, people often like to write uncivil words in spellings that are a derivative form. For example, people often use *sh!t* instead of *shit*, which clearly are the same thing in a public discourse. Hundreds of these variations may exist, making for a challenging identification problem.

Another challenge in identifying incivilities is that people can be really creative when they try to attack someone. This often happens when someone tries to indulge in ad hominem attacks with plausible deniability. For example, we have observed people using the word *DemocRat* instead of *Democrat* to identify someone with a democratic political orientation. Although these two words look similar, and sound exactly the same, *DemocRat* indicates that the target democrat is also a *rat*, a colloquial word for a spy, or a dishonest person. There are many other examples of this kind of variation, e.g. *democraps*. This phenomenon is sometimes referred as *Obscenity Obfuscation*, and researchers have found that it is becoming increasingly common in user generated contents in all sorts of social media platforms (Rojas-Galeano, 2017).

5.3 Difficulty in Comprehension

It is sometimes difficult to understand whether a word or a phase is used in an uncivil manner without understanding the context. For example, the word *lazy* can be used to describe the state of something that is actually slow or ineffective (e.g., *lazy algorithms*), or it can be used as an ad hominem attack on someone (e.g., *the lazy politicians have ruined this country*). As understanding the context of a content in a public discourse is difficult, separating these cases based on their contexts is challenging.

6 Incivility Prediction

In this section, we focus on our attempt to create a machine learning model that can be used as an incivility filter for moderators in social media platforms. Our model will exclusively use features obtained from the contents and reciprocations in the platform, while avoiding the demographic information that was used heavily by prior work. This will allow our models to be used on the large portion of online discourse where such demographic information is unavailable, e.g., where users are anonymous.

6.1 Data preparation

We will train our incivility prediction models on the Coe et al. (2014) data discussed in section 4. However, that data were designed for use in social science research, not natural language processing research, and thus there were several challenges in working with the data as they were collected, including:

- The comments were saved in PDFs, and the metadata referenced each comment by a number that was drawn (not typed) into the PDF beside the comment.
- The naming conventions for the files were inconsistent (spelling variations, variable length identifiers, etc.)
- Dates were saved using multiple formats (ddmmyy, dd-mm-yy, etc.)
- There were no specific markers in the text that identified the start and end of a comment.
- Many comments contained quotations from other comments, also with no consistent markers of where quotes began or ended.

We solved these problems using a combination of regular expressions (e.g., for normalizing dates), brute-force techniques (e.g., quotations were identified by comparing against all previous comments), and manual revision (e.g., renaming the files whose names were too inconsistent to be resolved automatically).

The resulting set of annotated comments were saved in JSON format for further computational analysis. We ended up with 6175 comments from the original set of 6444 comments after the extraction and cleaning process.

6.2 Prediction Task

Our main focus was to build a prediction model that can work as a filter for incivility in public discourse. We were also interested in how a model trained on public discourse data would work on a social media platform. We first divided our dataset into three smaller sets: train, development and test sets. Comments are randomly assigned to sets, and we ended up with 3950 comments in the training set, 989 comments in the validation set and 1236 comments in the test set. We set the the test set aside for our final evaluation, and worked only on the training and validation dataset to find the best model that can fit the problem.

6.3 Baselines

We found a similar task in Kaggle[1] (Wulczyn et al., 2017) that tries to identify toxicity of comments in the discourse section of Wikipedia. In that task, the best performing model was a recurrent neural network model with gated recurrent units (GRUs; Cho et al., 2014), but some simple non-sequential models (logistic regressions and support vector machines) also performed almost as well as the sequential model on that task.

For our baseline, we used two non-sequential machine learning techniques: logistic regression and support vector machines, using TF-IDF vectors obtained over words in the comments. We also considered a state-of-the-art out-of-the-box text classification model as a baseline, the Flair text classification model (Akbik et al., 2018), which uses GloVe word embeddings (Pennington et al., 2014) and pre-trained contextual word embeddings derived from two character-level language models. Flair achieved state-of-the-art performance in part-of-speech tagging and named-entity recognition tasks, and we thought that the character-based nature of the Flair model might be helpful in the face of the linguistic variation and creativity challenges we discussed earlier.

6.4 Model

Our model was inspired by the top performing systems in the Kaggle competition, and started with FastText embeddings (Joulin et al., 2016) for each of the words in a comment. These word vectors were fed to a recurrent layer consisting of bidirectional GRUs. The outputs of the GRUs were fed to an average pooling layer and a max pooling layer, which were then concatenated[2]. The output of the pooling was then fed through a sigmoid layer to produce the outputs. To avoid overfitting, we used a dropout layer (Srivastava et al., 2014) with 0.2 probability in between the input and hidden layer. We set the maximum length of input to 500 words for each comment, as this garnered the best validation performance in our preliminary analysis. We set class weights based on the frequency of name-calling and vulgarity: non-name-calling comments are 7 times more common than the name-calling ones, and non-vulgar comments are 35 times more common than vulgar ones, so we used a weighting scheme of 1:7 for name-calling and 1:35 for vulgarity. The model was trained with the Adam optimizer (Kingma and Ba, 2015) on mini-batches of size 32, with other hyperparameters set to their defaults. We trained each instance of this model for at most 500 epochs, with the option of early stopping if the validation accuracy did not improve for 10 consecutive epochs. A general structure of this model is shown in figure 1.

To further improve our model, we wanted to incorporate any metadata that were available to use. Coe et al. (2014) found that the thumbs up and thumbs downs received by a comment, the section of the article, and the author of the article all had some significance regarding incivility in the forum. So we introduced these metadata as features in our model. We created normalized feature vectors built on these attributes, and introduced them as auxiliary features right before the sigmoid layer, by concatenating them with the output of the pooling layers.

We also explored external resources that could improve our model. We created a pretrained model on the Kaggle dataset discussed earlier, as it had a large amount of annotated comments (over 160 thousand comments obtained from Wikipedia contributor's community). We used the same RNN model to train on the Kaggle data until it reached convergence, then retrained the model using our *Arizona Daily Star* data. The only portion of the model that was not shared between the pre-training (on Kaggle) and the training (on Arizona Daily Star) was the output sigmoid layers.

6.5 Experimental Results

The performance of the different models can be seen in table 1. Flair outperformed both of the other two baselines (36.55 vs. 23.35 and 18.46 F_1 in name-calling. Logistic regression and support vector machine models failed to detect single

[1] `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`

[2] This type of pooling worked well for Demidov (2018), and also performed well in our preliminary analysis.

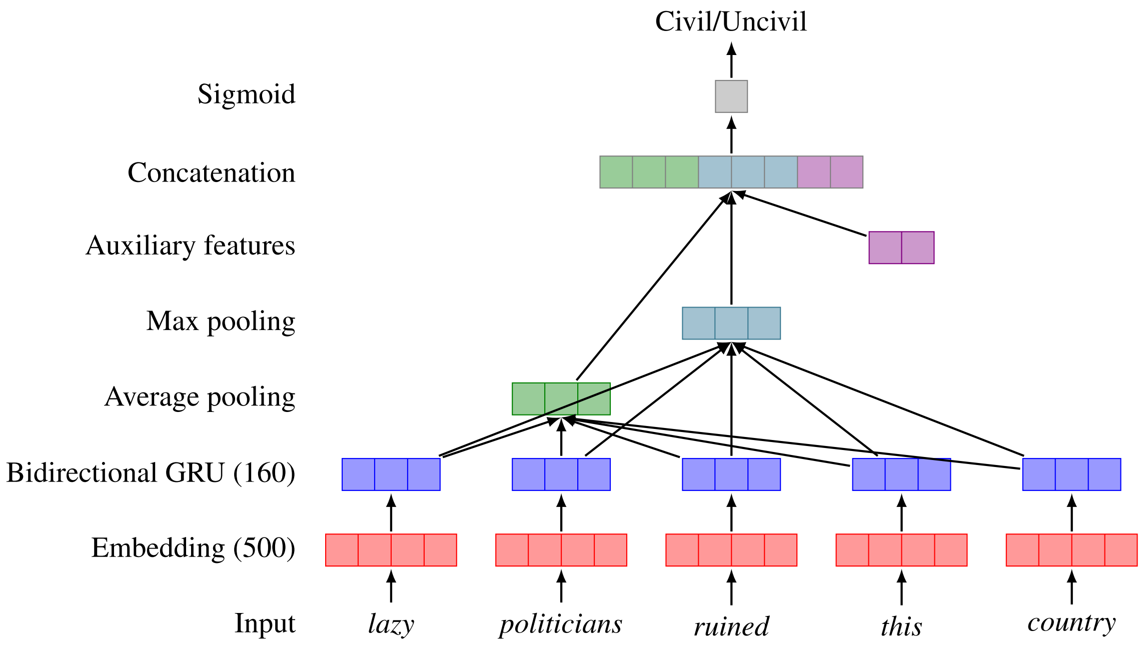

Figure 1: General structure of the RNN model. Auxiliary features are optional.

Validation						
	Name-calling			Vulgarity		
	Prec	Rec	F_1	Prec	Rec	F_1
Logistic regression	56.13	11.05	18.46	-	0.00	0.00
Support vector machine	54.10	14.89	23.35	-	0.00	0.00
Flair	52.17	28.12	36.55	25.00	7.41	11.43
GRU	43.65	61.72	51.13	37.50	66.67	48.00
GRU with auxiliary features	44.38	59.85	50.96	37.50	66.67	48.00
GRU with pretraining	69.44	19.53	29.79	50.00	11.11	18.03

Test						
	Name-calling			Vulgarity		
	Prec	Rec	F_1	Prec	Rec	F_1
GRU	45.76	50.63	48.07	48.72	57.57	52.77

Table 1: Performance of the models in terms of Precision (Prec), Recall (Rec), and F-measure (F_1).

instances of vulgarity in the development dataset, hence, Flair automatically outperformed these two. But our GRU-based model easily outperformed the Flair model (51.13 vs. 36.55 F_1 in name-calling, and 48.00 vs. 11.43 F_1 in vulgarity). These results stand in contrast to the Kaggle competition on toxicity detection, where such baselines performed nearly as well as the best (GRU-based) model, and all models achieved high levels of performance (>0.98 area under receiver operating characteristic curve). This suggests that the finer-grained incivility detection formulated by Coe et al. (2014) is more challenging than simple toxicity detection.

Adding the auxiliary features (upvotes, etc.) to the GRU-based model had virtually zero effect, with slight improvement on the model's precision but a slight drop in recall for name-calling, and absolutely no change for vulgarity. Using the Kaggle dataset to pre-train our GRU-based model before training on the *Arizona Daily Star* data yielded very high precisions, but at the cost of very low recalls. This suggests that while there is some overlap between the two tasks (toxicity detection and incivility detection), the differences between the tasks make it difficult to directly leverage the data from one task in the other.

name-calling		Vulgarity	
Tweet text	Score	Tweet text	Score
RT Jason_toronto: immigrant4trump Delusional Waters, Head Clown Schumer, Joke Perez, Senile Pelosi, Sleazy Schiff	0.997	Damn #BillCosby !! Damn damn damnnnn	0.996
#IHateItWhen incompetent idiots try to teach us how to live	0.997	I'm just going to say it. This is the stupidest tweet I've seen today. This BS bullying is not	0.973
@dapsixer GOP POTUS GOPChairwoman Primary these GOP candidates	0.979	”White Nationalism” WTH came up with this moniker? democrats?	0.985
#alis Dobbs obliterates Mitch McConnell and his pathetic excuses	0.989	Hell hath no fury like a bureaucrat scorned	0.969

Table 2: Examples of the GRU-based model predictions on the Russian troll Twitter data.

Since the GRU model with no auxiliary features or pre-training performed best on the development set, we evaluated the performance of this model on the test set. It achieved 48.07 F-measure for name-calling and 52.77 for vulgarity, scores roughly similar to what we had seen on the development data.

7 Incivility Prediction in Twitter

Though we built our models to detect incivilities in newspaper comments, we were interested in how well they would perform in other domains of social media. Karan and Šnajder (2018) has showed that cross-domain adaptation for detecting abusive language is possible- hence we would like to observe how well our model performs on a set of tweets.

In June 2018, The United States House Intelligence Committee released a list of 3841 Twitter account names that were human-operated troll accounts associated with Russia's Internet Research Agency. Darren Linvill and Patrick Warren from Clemson University collected all the tweets published since June 2015 from these accounts, cleaned them, and published a set of almost 3 million tweets (Linvill and Warren, 2018). These tweets are publicly available in FiveThirtyEight's Github page[3].

As prior research suggest that trolls are a big source of incivility in social media platforms (Fauman, 2008; Hinduja and Patchin, 2008), we took this opportunity to observe how our model performs on this dataset. We downloaded all the tweet texts and ran our GRU-based model on these texts. Results of this experiment can be found in the author's GitHub repository[4].

7.1 Observations

Our model identified 13% of all tweets as name-calling and 1.7% as vulgarity. These are roughly similar to the Arizona Daily Star training data, which had 14% name-calling and and 2.8% vulgarity. Though we do not have access to the expert annotators used by Coe et al. (2014), but we can nonetheless get an approximate measure of our model's performance by sampling predictions from our model and estimating the true label following the Coe et al. (2014) annotation guidelines.

To measure our model's precision, we took the 250 tweets that our model was most certain contained name-calling, and the 250 tweets that our model was most certain contained vulgarity. We manually reviewed each of these 500 tweets, and found only 7 instances of mistakenly tagged name-calling and 5 instances of mistakenly tagged vulgarity. To get a rough sense of our model's recall, we looked at the other end of the model's prediction spectrum. Based on a manual review of the model's prediction, the model almost never makes a mistake when the prediction score is below 10%; we found only one instance of mistaken name-calling, and no instance of mistaken vulgarity in the bottom 250 tweets that we manually annotated.

Table 2 shows some example tweets and the prediction scores from our model. The bottom two examples under name-calling and the bottom one example under vulgarity represent mistakes. In the first name-calling error, the model is confident (probability 0.979) that there is a name-calling,

[3] https://github.com/fivethirtyeight/russian-troll-tweets

[4] anonymized

perhaps because the terms GOP and POTUS frequently appear with name-calling in our training data. In the second name-calling error, the model is confident (probability 0.989) that there is a name-calling, likely because of the presence of the word *pathetic*, which is an aspersion, attacking an idea, not a name-calling, attacking a person. In the vulgarity error, *hell* has not been used to reference the religious concept of hell, but the word strongly associated with vulgarity in the training data. The table also shows some examples of reasonable successes of the model, for example, handling vulgar abbreviations like BS (short for *bullshit*) and WTH (short for *Who the hell*).

8 Future Works and Conclusion

Our work here aims towards keeping a civil environment in public discourse forums and social media platforms. Our goal was to build a filtering system that could work alongside human moderators to reduce their workload, be objective and independent of user reporting, and perform well on previously unseen social media streams. There is much work to do in this area: annotation of a large random sample of the troll tweets can give a more thorough estimate of model performance, and various forms of domain adaptation like self-training might be applied to improve the performance of the model. We have used word n-grams for features in our baseline models, which can be improved by using features obtained from domain-specific lexicons. There are lexicons of abusive words (Wiegand et al., 2018)- which can be used to create non-sequential models with smaller feature sets. Whether these simpler models are better is yet to be proven - as Agrawal and Awekar (2018) has shown that vocabulary of words used for cyberbullying varies significantly from one social media platform to another. They have also showed that swear words are not necessary to be uncivil in online social media- hence these types of detection techniques should not rely on such hand-crafted features.

One research question that follows this work is to observe whether incivility affects user engagement in social media. Prior research has observed that receiving replies can have effects in a user's engagement (Joyce and Kraut, 2006; Sadeque et al., 2015), and the language of these replies can also have consequences (Arguello et al., 2006). Habernal et al. (2018) has showed that 48% of comments that included ad hominem attacks ended the argument – which is indicative of lower engagement by the entire community. Hence, we believe that incivility has a significant influence on user engagement, and in turn may contribute to a community's sustainability. This is yet to be proven, and more work needs to be performed to prove or disprove this hypothesis.

In this paper, we have presented a recurrent neural that can identify incivilities in public discourse. Though trained on a corpus of newspaper comments, we have initial evidence that it also performs well in detecting incivilities in Twitter. We believe our model will be able to serve as a wide-range incivility filter in other social media platforms.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. *CoRR*, abs/1801.06482.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 959–968, New York, NY, USA. ACM.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

Vladimir Demidov. 2018. Kernel submission for kaggle toxic classification challenge. `https://www.kaggle.com/yekenot/pooled-gru-fasttext?` Last Accessed: 2018-12-02.

Michael A. Fauman. 2008. Cyber bullying: Bullying in the digital age. *American Journal of Psychiatry*, 165(6):780–781.

Michael Gauthier, Adrien Guille, Fabien Rico, and Anthony Deseille. 2015. Text mining and twitter to analyze british swearing habits. In *Handbook of Twitter for Research*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv preprint arXiv:1802.06613*.

Sameer Hinduja and Justin W. Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, 29(2):129–156.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Elisabeth Joyce and Robert E. Kraut. 2006. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Kate Kenski, Kevin Coe, and Stephen A. Rains. 2017. Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, 0(0):0093650217699933.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. International Conference on Learning Representation.

Darren L. Linvill and Patrcik L. Warren. 2018. Troll factories: The internet research agency and state-sponsored agenda building. `https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-\Research-Agency-and-State-Sponsored\-Agenda-Building`.

Justin W. Patchin and Sameer Hinduja. 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2):148–169.

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

Stephen A. Rains, Kate Kenski, Kevin Coe, and Jake Harwood. 2017. Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, 22(4):163–178.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Sergio Rojas-Galeano. 2017. On obstructing obscenity obfuscation. *ACM Trans. Web*, 11(2):12:1–12:24.

Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. 2015. Predicting continued participation in online health forums. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 12.

Weber Shandwick. 2018. Civility in america 2018: Civility at work and in our public squares. `https://www.webershandwick.com/wp-content/uploads/2018/06/Civility-in-America-VII-FINAL.pdf`. Last Accessed: 2018-06-11.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '14, pages 415–425, New York, NY, USA. ACM.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. 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 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. 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.