Appendix A

Sub-reddit Annotation

Seven coders received initial training on the coding procedures, and each of them completed five iterative pilot coding exercises. Coders were instructed to code sub-reddits based on the self-provided sub-reddit description and the top popular posts in that sub-reddit. More specifically, coders would go to the webpage of the sub-reddit and read “about community” for an initial sense of the sub-reddit. Then, they would read the top 20 posts on the first page. Although the coders only had access to the posts and comments from 2019; based on our knowledge about Reddit dynamics, sub-reddits are rather stable in terms of discussion themes even though it is possible that sub-reddits develop and change over time to some extent. If a sub-reddit was active in 2019, the whole discussion in that sub-reddit is identifiable via top posts and comments in 2019. If a sub-reddit was “dead,” the top posts from the last activity served the purpose of identifying the discussion topic. If 70% of posts were related to socio-political issues, the sub-reddit was categorized as a political sub-reddit. Then, the coders were instructed to judge by the top posts whether the sub-reddit is homogeneous or heterogeneous, and its political leaning if the sub-reddit is politically homogeneous. If less than 40% of the posts were related to socio-political issues, the sub-reddit would be categorized as non-political, and coders would further categorize it based on its domain topic. The rest sub-reddits would be categorized into the mixed group. Mixed sub-reddits would be further categorized as politically homogeneous/heterogeneous and for their political leaning, as well as different non-political topics. Coding procedure took place from October 2019 to March 2020. The inter-coder reliability for sub-reddit types of political/non-political/mixed (Fleiss’s kappa=0.62), political homogeneous/heterogeneous (Fleiss’s kappa=0.59), mixed homogeneous/heterogeneous (Fleiss’s kappa=0.85), political liberal/conservative (Fleiss’s kappa=0.76), mixed liberal/conservative (Fleiss’s kappa=0.59), non-political topics for mixed group (Fleiss’s kappa=0.57), and non-political topics (Fleiss’s kappa=0.83) were then calculated and met the fair requirement. These values of Fleiss’ Kappas are fair and acceptable, especially given that we had 7 coders, instructed to code communities into 10 categories, and because each community could be categorized into multiple categories. Coders then proceeded to individual coding, with 10% overlap in the coded content.

Appendix B

**Table 1**

*Example Sub-reddits for Each Category*

|  |  |  |
| --- | --- | --- |
| Political | Homogeneous liberal | r/neoliberal, r/Libertarian, r/nottheonion, r/changemyview |
| Homogeneous conservative | r/The\_Donald, r/Conservative, r/Firearms, r/forwardsfromgrandma |
| Heterogeneous | r/politics, r/news, r/bestof, r/Vaping |
| Mixed | Homogeneous liberal | r/atheism, r/NoStupidQuestions, r/conspiracy, r/technology |
| Homogeneous conservative | r/BlackPeopleTwitter, r/terriblefacebookmemes, r/ar15, r/CCW |
| Heterogeneous | r/AskReddit, r/unpopularopinion, r/MGTOW,r/wallstreetbets |
| Non-political | Gaming | r/FortNiteBR, r/gaming, r/leagueoflegends, r/DestinyTheGame |
| Entertainment | r/funny, r/movies, r/videos, r/videos, r/anime |
| Sports | r/nba, r/nfl, r/soccer, r/SquaredCircle |
| Health | r/Drugs, r/depression, r/opiates, r/fatlogic |
| Music | r/hiphopheads, r/Music, r/popheads, r/kpop |
| Science/Technology | r/pcmasterrace, r/buildapc, r/Android, r/apple |
| Pets/Animals | r/aww, r/cats, r/NatureIsFuckingLit, r/dogs |
| Lifestyple/Fashion | r/gonewild, r/relationships, r/mildlyinteresting, r/CryptoCurrency |
| Other | r/todayilearned, r/AskOuija, r/Showerthoughts, r/WTF  |
| Memes | r/dankmemes, r/MemeEconomy, r/me\_irl, r/PewdiepieSubmissions |

Appendix C

Incivility Classifier Building

BERT (Bidirectional Encoder Representation from Transformers) is a deep transformer model pre-trained on huge amounts of unlabeled text using a word-masking training objective. Word-masking training objective refers to sequences of words where 15% of the words were replaced by a mask. The machine learns the language model by predicting the actual words under the mask based on the context provided from un-masked words in the sequence. By training on a word-masking objective with a large amount of unlabeled data, BERT builds a powerful language model which can then be fine-tuned to successfully complete specific language-related tasks with its knowledge of the language, such as question answering and text classification. Fine-tuning refers to strategies that take weights of a pretrained model to use as an initialization to build a more specific model on labeled data from the same domain.

DistilBERT is a BERT-based model which has had its number of parameters reduced using a compression technique called model distillation, in which a smaller model - the student - is trained to replicate the output of a larger model - the teacher - using the teacher-student training technique (Sahn et al., 2019). The result is a smaller, faster student model with the same architecture as the teacher model but fewer layers, which retains much of the performance of the teacher model. According to Sahn et al. (2019), the size of a BERT model can be reduced by 40%, while 97% of its language understanding capabilities can be retained; additionally, it can be up to 60% faster.

Our models started with the respective uncased “base” pretrained language models from HuggingFace’s Transformer’s package. The models were then further pretrained on a dataset of 3-million unlabeled Reddit comments sampled from the whole data with 100,000 training steps using a masked language modeling task (as suggested in Sun, et al., 2019. Training steps refers to the number of iterations, in other words, it is the number of batches that were processed. In our case, 100,000 batches have been processed during training, each batch has 30,000 comments. Masked language modeling is a task in which words in the input data are replaced at random with a “mask” token; the model’s job is to recreate the original text from this masked input). Because off-the-shelf BERT models are trained on general web-scraped data, their language models may not be well-suited to the domain of a target task. This domain pretraining procedure promotes the adaptation of the language model to the target domain, in our case Reddit posts. Once pretraining is complete, our models were fine-tuned for four epochs (that is, four iterations through the dataset) on 5000 annotated comments, with 10% of the data set aside for training validation and 1000 coded comments set aside for model testing (see Davidson et al., 2020 for details). Classification results using BERT and DistilBERT pretrained models are shown in Table 1.

Despite their promising performance, our BERT and DistilBERT-based models are too computationally expensive and time-consuming to apply to the entire 13-year Reddit dataset. Our approach to deal with data at this scale is to generate a large collection of machine-labeled comments from our unlabeled dataset using the fine-tuned DistilBERT model; we use these machine-labeled examples as training data for a more computationally efficient model, namely a logistic regression binary classifier. Considering that the natural proportion of incivility in Reddit comments may lead to an imbalanced training set and potential of bias in models, an additional synthetic data for oversampling was generated using ADASYN (He et al., 2008, an algorithm to generate synthetic data with more “hard to learn” examples). In total, we generated 5 million Reddit comments using our DistilBERT model. Logistic regression models were then trained on this synthetic data together with a much smaller human-coded dataset (see Davidson et al., 2020). Classification results are shown in Table 2.

To evaluate our models, we use Precision, Recall, and F1 score, all of which are standard methods of evaluating classification models. Precision is simply the ratio of correct predictions for each class to the number of total predictions for that class. Recall, by contrast, is the ratio of correct predictions for each class to the total number of items which should belong to that class. This, Precision tells you how many of your predictions were correct, while Recall tells you how many of the items belonging to the target class your model caught. F1 score is simply the harmonic mean of Precision and Recall. Because we wish to ensure that our model is predicting the less common “positive” class accurately, we use macro-averaged F1, which weights the positive and negative classes equally in the F1 calculation, regardless of the respective size of each class.

**Table 1.** *Classification Results from BERT and DistilBERT Models*

|  |  |  |
| --- | --- | --- |
|   | BERT | DistilBERT |
| precision | 0.814 | 0.936 |
| recall | 0.759 | 0.702 |
| F1 | 0.784 | 0.802 |
| accuracy | 0.956 |  |

**Table 2.** *Classification Results for Different Training Data*

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data | Precision | Recall | *F1* |
| Human coded data | 1 | 0.173 | 0.295 |
| Synthetic data | 0.835 | 0.731 | 0.779 |

Appendix D

**Table 1.**

*Annotation and Classification Examples.*

|  |  |
| --- | --- |
| Human Annotated Uncivil Comments | Hey fag, go kill yourself |
| I lie on reddit all the time. You fuckers will never catch me! |
| good on you buddy hope you made a difference to that young lass, unfortunately the world is full of wankers and we all find different coping mechanisms you have my upvote for being awesome!! |
| YOU CANT PROVE THAT, YOR GOING TO HELL! STOP BELITTLING MY BLEEFS!! |
| Human Annotated Civil Comments | You guys have obviously never seen a malfunctioning weather balloon traped in a pocket of super heated swamp gas. |
| But I doubt any of them would say it wasn't worth it. |
| Meh, vodka is always a good thing. |
| Sometimes I think people just create stories to get attention. |
| Machine Classified Uncivil Comments | Is… Is this on fucking LinkedIn? I cannot despise that shit service more  buit you're pushing it OP |
| It's pretty trash on all platforms frame rate wise. |
| This is beyond pathetic - Trump is a PSYCHOPATH. We must vote him in out in 2020. VOTE BLUE NO  MATTER WHO! |
| Murder victims aren't fucking blamed for their own murder unlike rape and sexual assault victimes. By the way, I never saud murder was fine so get off your fucking high hoprse bitch. |
| Machine Classified civil Comments | It was compulsive |
| I just had dinner, but definitely have room for this delicious dessert ;) |
| I know a lot of people who wished they'd spent more time with family, friends, or living life. I don't know anyone who has looked back and wished they'd just put in more hours for more money. |
| No, but still more than the US. |

Appendix E

As additional exploratory analysis, we also looked at incivility trends in non-US English sub-reddits. In total, we identified 189 sub-reddits (2.23% of total identified sub-reddits) that explicitly discuss geolocations or topics outside of the US. Among which, 26 sub-reddits are political, 51 are mixed, and the remaining 112 are non-political sub-reddits. The average overall incivility in the non-US group ranged from 11 to 13%, similar to the over-time trend in the US group with a slight decrease before 2015 reaching the lowest point at 11.22% and returning to a level around 12%. In the same manner, the incivility proportion in *political* groups was higher than overall trend (*t* = 6.684, *df* = 11, *p* < 0.001), *mixed* groups (not significantly, t = -2.210, df = 11, p = 0.058) and *non-political* groups (*t* = 19.21, *df* =12, *p* < 0.001)). As the majority of non-US English *political* sub-reddits were about UK and Europe, the steady increase of incivility in those sub-reddits in 2016-2017 may respond to Brexit. The temporal variations in incivility in *mixed* groups, is higher than that in *non-political* groups (*t* = 4.619, *df* = 11, *p* < 0.01) and - at some points, such as during the 2016-2017 Brexit period - was similar to that in political groups, ranging between 8 and 13%. Similar to the incivility trend in *non-political* groups in the US sub-reddits, *non-political* groups in non-US sub-reddits contained the least level of incivility ranging between 8 to 10%.

Appendix F

**Table 1**

*Average yearly Incivility Proportion for All Categories from 2008 to 2019*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Political | Mixed | Non-political | Total  |
| Heterogeneous | 14.46% | 12.07% |  |  |
| Homogeneous | 13.92% | 11.48% |  |  |
| Liberal | 13.61% | 11.38% |  |  |
| Conservative | 13.35% | 12.16% |  |  |
| Gaming |  | 15.36% | 8.25% |  |
| Entertainment |  | 10.17% | 10.57% |  |
| Sports |  | 13.75% | 11.57% |  |
| Health |  | 9.89% | 11.78% |  |
| Music |  |  | 10.46% |  |
| Science/Technology | 9.10% | 5.40% |  |
| Pets/Animals |  |  | 6.24% |  |
| Lifestyle/Fashion |  | 12.20% | 9.50% |  |
| Memes |  | 14.24% | 10.95% |  |
| Others |  | 11.56% | 11.32% |  |
| Total  | 14.12% | 11.89% | 10.05% | 10.68% |