Developing a New Classifier for Automated Identification of Incivility in Social Media

Sam Davidson[*], **Qiusi Sun**[†], **Magdalena Wojcieszak**[†‡]
[*]Dept. of Linguistics, [†] Dept. of Communication
University of California, Davis
[‡]University of Amsterdam
{ssdavidson, qssun, mwojcieszak}@ucdavis.edu

Abstract

Incivility is not only prevalent on online social media platforms, but also has concrete effects on individual users, online groups, the platforms themselves, and the society at large. Given the prevalence and effects of online incivility, and the challenges involved in human-based incivility detection, it is urgent to develop validated and versatile automatic approaches to identifying uncivil posts and comments. This project advances both a neural, BERT-based classifier as well as a logistic regression classifier to identify uncivil comments. The classifier is trained on a dataset of Reddit posts, which are annotated for incivility, and further expanded using a combination of labeled data from Reddit and Twitter. Our best performing model achieves an F_1 of 0.802 on our Reddit test set. The final model is not only applicable across social media platforms and their distinct data structures, but also computationally versatile, and - as such - ready to be used on vast volumes of online data. All trained models and annotated data are made available to the research community.

1 Introduction

Given the growing polarization in the United States, the increasing popularity of partisan media, and the widespread use of social media for information and discussion (see Iyengar et al. (2019) for a review), many scholars and observers worry about the accelerated use and spread of incivility in the online environment. Incivility, defined as "features of discussion that convey disrespectful tone toward the discussion forum, its participants, or its topics" (Coe et al., 2014) is a common aspect of many online communities, especially anonymous forums (Reader, 2012) such as Reddit. Estimates suggest that more than 84% of Americans have experienced incivility online, and among those who have ever experienced it, the number of average weekly encounters with incivility was as high as 10.6 times (KRC Research, 2018). In addition to lowering the standards of public discourse, incivility has concrete effects on users, online discussions, and social media platforms. The use of and exposure to incivility generates negative emotions, such as anger, anxiety, or mental distress, and is related to aggression (Gervais, 2015) and hostile communication (Groshek and Cutino, 2016). Incivility also turns users away from online discussions altogether (Anderson et al., 2014; Bauman et al., 2013; Moor et al., 2010; Ransbotham et al., 2016). Given these reasons for the public and industry to be concerned with online incivility, many companies seek to automatically detect incivility in order to understand its scope, identify the online communities in which incivility is particularly prevalent, and - ultimately - address the problem.

This project offers a step in this direction. We present machine learning models for detecting incivility in social media, models that are not only computationally efficient but also applicable across platforms. We propose both a BERT-based neural classifier as well as a logistic regression based classifier trained on manually annotated and artificially labeled data. Our results suggest that the proposed models perform well across distinct data/communication structures of different platforms, and, as such, can be easily applied to detect incivility.

2 Previous Work

There is considerable conceptual and operational ambiguity in the literature on incivility and related concepts under the umbrella of offensive or intolerant speech (see (Rossini, 2020) for a review). Some studies use incivility interchangeably with hate speech, which refers to speech that aims to discriminate against a certain identity group, or

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 95–101

https://doi.org/10.18653/v1/P17

aggressive or toxic language, which includes personal attacks (Rösner and Krämer, 2016). However, incivility is a broader concept, which focuses on content that goes against acceptable social norms in terms of vulgarity, name-calling, or offensive language (Papacharissi, 2004), whereas hate speech or aggressive language captures more specifically discourse that offends, derogates, or silences others and may promote harm (Rossini, 2020). Increasingly, incivility is conceptually and operationally distinguished from such intolerant discourse, and evidence suggests that the effects of these two forms of expressions also differ (Rossini, 2020). Definitions of incivility vary, ranging from "a norm-defying behavior" (Gervais, 2015), "an explicit attack" (Anderson and Huntington, 2017), to the violation of interpersonal politeness norms (Mutz, 2015; Mutz and Reeves, 2005), yet most include a lack of respect toward discussion participants or arguments (Santana, 2014), and a impolite tone of discourse (Papacharissi, 2004). The often used definition, which we adopt for the purpose of our machine learning model, sees incivility as features of discussion that convey disrespectful tone toward the discussion participants or its topics, including name-calling, mean-spirited or disparaging words directed at a person or group of people, an idea, plan, policy, or behavior, vulgarity, using profanity or language that would not be considered proper in professional discourse, and pejorative remarks about the way in which a person communicates (Coe et al., 2014). As such, our approach encompasses both the less societally detrimental foul language or harsh tone as well as the more intolerant discourse.

From a technical perspective, previous research using machine learning models to detect incivility and other offensive or intolerant language online has focused primarily on the use of logistic regression (Theocharis et al., 2020; Daxenberger et al., 2018; Maity et al., 2018), support vector machines (Joksimovic et al., 2019; Maity et al., 2018), and various neural classification models (Sadeque et al., 2019). BERT (Devlin et al., 2019) and related transfomer language models have been used in related tasks, such as identifying abusive language on Twitter (Nikolov and Radivchev, 2019; Risch et al., 2019), including many entrants in the OffensEval task at SemEval-2020 (Zampieri et al., 2020). To our knowledge, this paper is the first to utilize a fine-tuned BERT model to identify incivility on social media platforms, and one of few projects that train the classifier on data from more than one platform. Also, past work on identifying incivility over time has mostly analyzed Twitter data during certain political events, such as the 2016 presidential election in the US (Siegel et al., 2018), and/or looked at political incivility in specific contexts (e.g., among politicians, e.g., (Theocharis et al., 2020). These rather narrow, single-platform foci limit the applicability of the developed classifications, a limitation we address in this project.

In addition to these contributions of our work, our primary contribution may lie in our data augmentation method. Specifically, we extend recent approaches to automatically label additional training data to improve the performance of a logistic regression classifier. Previous work in detection of offensive language has used back-translation (Ibrahim et al., 2020) and data transformation techniques (Rizos et al., 2019) to augment limited training data. While some work (Theocharis et al., 2020) utilizes the Google Perspectives API to label additional training data, which introduces noise to the operationalization of incivility, we take advantage of our well-performing BERT classification model to generate artificial training data for a logistic regression classifier. The resulting classifier can be efficiently run on CPU and is far less computationally expensive than our comparably performing BERT model. This extension makes our classifier easily applicable to vast amounts of data and readily implemented on social media platforms or the comments sections of websites of news media organizations.

3 Dataset

We present a corpus of Reddit data annotated for incivility at the post level, and further annotated for political/non-political content at the subreddit level. We chose to use Reddit as the source of our dataset because Reddit is the sixth most popular website in the US and the third most visited social media platform following YouTube and Facebook (Alexa.com, 2019). There are more than 430 million active users worldwide on Reddit (Perez, 2019). Also, the anonymous nature of Reddit makes it popular for sharing information and engaging in long and complex discussions, making it ideal for observing online discourse. Furthermore, the public nature of Reddit allowed us to gather a large number of posts from across various user

communities, known as subreddits. Each subreddit has a general topic, behavioral norms, and community standards, allowing for a creation of a diverse dataset, which further increases the applicability of the resulting machine learning model.

To tackle the detection problem, we identified the most popular subreddits from 2006 to 2019 that contained 95% of the total comments by (1) the number of comments in the subreddit each year, and (2) the number of followers that commented in the subreddit each year, which resulted in 9355 subreddits across the years. We then collected 5000 comments from these subreddits using stratified random sampling technique, such that the random sampling from each year is based on each year's proportion in the total number of comments. These 5000 posts were the manually labeled.

3.1 Dataset Annotation

Instead of adapting annotation schema that focused on profanity and swear words or phrases (i.e., the more narrow definition of incivility) (Zampieri et al., 2019; Mohan et al., 2017; Almerekhi et al., 2020), we developed a coding manual to classify comments according to four dimensions present in offensive speech more broadly. We account for whether a comment contains: (1) name-calling, mean-spirited or disparaging words directed at a person or a group of people; (2) aspersion, mean-spirited or disparaging words directed at an idea, plan, policy or behavior; (3) pejorative or disparaging remark about the way in which a person communicates, and (4) vulgarity, profanity or language that would not be considered proper. Our operational approach accounted for the content aspect (e.g., vulgarity or profanity, such as "you're a dumbass for simplifying the issue and trying to jump right into the helm of the 'y'r all hypocrites' bandwagon") and the different targets of incivility or foul content included in the intolerant discourse (e.g., "... the interests of left-handed black female dwarves"), to create a comprehensive and inclusive annotated dataset for model building. Annotators were asked to apply a binary label to indicate whether or not the comment contains incivility. The annotators were three undergraduate students in social sciences at UC Davis, two native English speakers and one with English as the second language. Two annotators are heavy Reddit users and one is a user of other social media. The annotators were trained on the definitions and procedures, and each of them completed five pilot coding exercises. Each annotator first independently coded a random set of 50 comments with Fleiss's kappa of 0.618. They then compared results, discussed and resolved discrepancies, and clarified confusions. These steps were repeated multiple times with increasingly large comment sets until an acceptable agreement level was reached. In total, all three annotators completed 1000 comments together during training, with Fleiss's kappa of 0.663. The major discrepancies pertain to potentially sarcastic comments (e.g., "Great, now we're paying for CBC to promote cuckoldry"), which some coders saw as uncivil and others as innocent sarcasm. After an acceptable coding precision was established among the three annotators, the remaining 4000 comments were randomly divided into three sets and each annotator independently coded an assigned set. The final result of this process is a set of 5000 comments labeled for incivility. Additionally, our dataset includes coding at the subreddit level to identify subreddits that were political, non-political, or mixed (i.e., contained some political and some non-political content). This allows us to analyze the prevalence of incivility across different kinds of online discussions and across the political spectrum.

4 Classifier Training

To demonstrate the efficacy of our collected dataset, we use supervised machine learning to automatically identify uncivil Reddit posts. However, annotating a dataset large enough to train a state-of-the-art neural classifier from scratch is a costly and time-consuming undertaking. We experimented with several neural binary classifiers, with our best-performing models built on top of transformer-based language models, namely BERT (Devlin et al., 2019) and its relative, DistilBERT (Sanh et al., 2019). Past work has demonstrated that fine-tuning large, pre-trained language models, such as BERT and DistilBERT, is an effective method for creating a high-quality neural classifier with limited supervised training data. As described in Sun et al. (2019), we conduct additional pretraining of the BERT-base and DistilBERT-base models on a large collection of Reddit posts as in-domain data. Once pretrained, we fine-tune our models for classification on our annotated dataset of Reddit comments, which trained annotators classified with binary labels for incivility.

Finally, in an effort to extend past work by creating a more flexible, platform agnostic classifier, we train a logistic regression classifier for incivility prediction in social media by combining the data presented in Theocharis et al. (2020) with our annotated and artificially labeled datasets.

Our Reddit dataset (including annotation disagreements), test predictions, scripts and models are available on the project GitHub repository [1].

5 Experiments and Results

Our BERT and DistilBERT models begin with the respective base pretrained language models, as implemented in HuggingFace's Transformer's package. We then further pretrain these models on dataset of 3 million Reddit posts, for 100,000 training steps (as suggested by Sun et al. (2019)) using the masked word prediction task (Devlin et al., 2019). We then utilize these pretrained models in a classification setup, utilizing a softmax layer to predict binary class probability based on the [CLS] token in BERT's final hidden layer. For classification fine-tuning, all inputs to the models are limited to 256 tokens in length, with a training batch size of 16. We use the AdamW optimizer (Gugger and Howard, 2018) with default learning rate and epsilon values. We fine-tune our model for classification for four epochs on our dataset of 5,000 Reddit posts which are coded for incivility, with 10% of the data set aside for training validation, and 1000 annotated posts set aside for model testing. Classification results using BERT and DistilBERT are shown in Table 1.

Model	Precision	Recall	F_1
BERT	0.814	0.76	0.786
DistilBERT	0.936	0.702	0.802

Table 1: Results - Models trained on Reddit data

One major goal of this project is to classify multiple years of Reddit data for further analysis of incivility across political and non-political subreddits. Despite the acceptable performance of our BERT classification models, the models were too computationally expensive to classify the approximately 800 million posts per year we collected from Reddit. To address this constraint, we also train a logistic regression classification model to be able to classify large amounts of Reddit data without the use of expensive neural classifiers. However, given the small size of our annotated training set, we must generate additional training data to train an effective logistic regression model. In order to improve system performance, we first use our fine-tuned DistilBERT model to classify a large collection of Reddit posts. We then uptrain a logistic regression model on this synthetic data, along with our annotated data. As detailed in Section 5, the resulting model achieves an F_1 score which is competitive with our BERT and DistilBERT models, while also being able to classify data more quickly and at lower computational cost, making our model widely applicable.

All logistic regression models are trained using TFIDF of stemmed unigrams as features. Given the relative imbalance of labels in our training data, in which positive examples of incivility represent only 10.3% of annotated posts, we use ADASYN (He et al., 2008) to generate additional synthetic data for oversampling. We train a second model on synthetic data consisting of 5 million Reddit posts which are labeled for incivility using our trained DistilBERT model. Results are shown in Table 2

To test the overlap of concepts such as hate speech and offensive language with incivility, we applied the classifier provided by Davidson et al. (2017) to the test portion of our Reddit dataset. To conduct our test, we combined the classes "offensive language" and "hate speech" predicted by the Davidson et al. (2017) classifier into a single class. On our Reddit data, this classifier achieves an F1 of 0.242, indicating limited overlap between the these domains. This test demonstrates that the definitional, conceptual, and operational differences between incivility and related domains of offensive speech are indeed represented in our labeled data.

In order to further test the efficacy of our implementation, we train a logistic regression model as outlined above using Twitter data collected and annotated by Theocharis et al. (2020). Finally, to create our platform agnostic model, we train a logistic regression model by combining our annotated and synthetic Reddit data with the annotated and synthetic data from Theocharis et al. (2020), which we test on the Theocharis et al. (2020) Twitter test set, as shown in Table 3.

6 Analysis and Discussion

Our encouraging results in classifying incivility in Reddit posts demonstrate the efficacy of our dataset

[1] https://github.com/ssdavidson/reddit_incivility

Training Data	Precision	Recall	F_1
Annotated Reddit	1	0.173	0.295
Synthetic Reddit	0.835	0.731	0.779
Reddit + Twitter	0.828	0.74	0.782

Table 2: Results - Logistic Regression Models on Reddit Test Data

Training Data	Precision	Recall	F_1
Synthetic Reddit	0.872	0.158	0.267
Reddit + Twitter	0.711	0.474	0.569

Table 3: Results - Logistic Regression Models on Twitter Data

for classifier training. We have applied our trained classifier to 95% of Reddit comments from the year 2017, finding that 9.21% of non-political comments are uncivil, compared to 14.75% of political comments; initial results that indicate relative prevalence of incivility in online *political* discourse.

Due to the scale of the data to be ultimately classified, we were concerned as much with computational efficiency as with prediction accuracy when building our incivility classifier. When we use our trained BERT model to generate a large quantity of synthetically labeled training data, the performance of our log regression model is comparable to that of the our fine-tuned BERT models.

Concern with computational efficiency also informed our choice of features in our logistic regression model. While alternate features could be used, such as Doc2Vec or Word2Vec embeddings, we chose to use TFIDF due to the simplicity of calculating these features. Additionally, the choice of TFIDF was informed by the work of Theocharis et al. (2020), who demonstrate the utility of TFIDF for the task for incivility classification. Finally, the fact that our TFIDF-based logistic regression model performs similarly well to the BERT model is evidence of the effectiveness of the choice of TFIDF features. That said, the use of alternate features may improve model performance, and we leave this to future work.

The similarity between the predictions made by our BERT model and our logistic regression model indicates that the logistic regression model retains much of the predictive power of the BERT model. In fact, across 996 test comments, the two models disagreed on only 27 comments, for a rate of 2.7%. From reviewing the disagreements we can identify several classes of comment on which the two models often disagree. The first, and most obvious, is very long comments. BERT is designed to truncate long input text (our implementation truncates inputs longer than 256 tokenized word pieces). Thus, our BERT model may mislabel longer comments in which the incivility occurs later in the comment. Another source of disagreement comes from the fact that our TFIDF-based classifier tends to be more sensitive to individual lexical items, which is to be expected as BERT is known to condense far more semantic information than do count-based vectorization techniques such as TFIDF (Jawahar et al., 2019). For example, our regression model mislabels the comment "This is dope! Does anyone know where I can purchase one for myself?" as an uncivil comment, presumably due to the presence of the word "dope", while our BERT model labels the comment correctly. In future work, we plan to conduct a more rigorous analysis of labelling disagreements between the two models to better understand the role of lexicon and compositional semantics in the incivility classification task.

Finally, we demonstrate the flexibility of our model training strategy by creating a combined incivility prediction model using our automatically labeled Reddit data with the synthetic data provided by Theocharis et al. (2020). The resulting model has shown promise as a platform agnostic incivility classifier model for social media.

7 Conclusion and Future Work

In this paper, we present a new dataset of Reddit posts annotated at the comment level for incivility, as well as at the subreddit level for political content. Further, we demonstrate the efficacy of this dataset to train machine learning models for incivility detection, both alone and in combination with previously available datasets, to create a platform agnostic classifier for incivility on social media.

Using our trained classifier, our future goal is to provide a systematic overview of trends in incivility on social media, across time and variety of discussion topics. The project aims to capture the fluctuations in the prevalence of incivility in political and non-political online spaces, politically homogeneous and heterogeneous discussions, liberal and conservative ones, and also among different non-political topics. The anticipated study will add our understanding of the development of online incivility and shed light on incivility interventions.

References

Alexa.com. 2019. Alexa - Top Sites in United States - Alexa.

Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators. In *Companion Proceedings of the Web Conference 2020*, pages 294–298.

Ashley A Anderson, Dominique Brossard, Dietram A Scheufele, Michael A Xenos, and Peter Ladwig. 2014. The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, 19(3):373–387.

Ashley A Anderson and Heidi E Huntington. 2017. Social media, science, and attack discourse: How Twitter discussions of climate change use sarcasm and incivility. *Science Communication*, 39(5):598–620.

Sheri Bauman, Russell B Toomey, and Jenny L Walker. 2013. Associations among bullying, cyberbullying, and suicide in high school students. *Journal of adolescence*, 36(2):341–350.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Johannes Daxenberger, Marc Ziegele, Iryna Gurevych, and Oliver Quiring. 2018. Automatically Detecting Incivility in Online Discussions of News Media. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 318–319. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bryan T Gervais. 2015. Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, 12(2):167–185.

Jacob Groshek and Chelsea Cutino. 2016. Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. *Social Media+ Society*, 2(4):2056305116677137.

Sylvain Gugger and Jeremy Howard. 2018. AdamW and Super-convergence is now the fastest way to train neural nets. https://www.fast.ai/2018/07/02/adam-weight-decay/.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. Alexu-backtranslation-tl at semeval-2020 task [12]: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22:129–146.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Srecko Joksimovic, Ryan S Baker, Jaclyn Ocumpaugh, Juan Miguel L Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated Identification of Verbally Abusive Behaviors in Online Discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45.

Powell Tate KRC Research, 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.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2018. Opinion conflicts: An effective route to detect incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of Reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Peter J Moor, Ard Heuvelman, and Ria Verleur. 2010. Flaming on YouTube. *Computers in human behavior*, 26(6):1536–1546.

Diana C Mutz. 2015. Incentivizing the manuscript-review system using REX. *PS, Political Science & Politics*, 48(S1):73.

Diana C Mutz and Byron Reeves. 2005. The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, pages 1–15.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-Radivchev at SemEval-2019 Task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

Sarah Perez. 2019. Reddit's monthly active user base grew 30% to reach 430M in 2019.

Sam Ransbotham, Robert G Fichman, Ram Gopal, and Alok Gupta. 2016. Special section introduction—ubiquitous IT and digital vulnerabilities. *Information Systems Research*, 27(4):834–847.

Bill Reader. 2012. Free press vs. free speech? The rhetoric of "civility" in regard to anonymous online comments. *Journalism & mass communication quarterly*, 89(3):495–513.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model. In *KONVENS*.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000.

Leonie Rösner and Nicole C Krämer. 2016. Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media+ Society*, 2(3):2056305116664220.

Patrícia Rossini. 2020. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, page 0093650220921314.

Farig Sadeque, Stephen Rains, Yotam Shmargad, Kate Kenski, Kevin Coe, and Steven Bethard. 2019. Incivility Detection in Online Comments. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 283–291.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Arthur D Santana. 2014. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice*, 8(1):18–33.

Alexandra A Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2018. Measuring the prevalence of online hate speech, with an application to the 2016 US election. https://smappnyu.wpcomstaging.com/wp-content/uploads/2018/11/Hate_Speech_2016_US_Election_Text.pdf.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. The Dynamics of Political Incivility on Twitter. *Sage Open*, 10(2):2158244020919447.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.