# Avoiding Dilution but Preserving Context: Fine-Tuning Incivility Classifier with Sentence-level Auxiliary Signals on Data from War-Related Subreddits

**Anonymous ACL submission**

## Abstract

This paper seeks to overcome current shortcomings in the literature by detecting content-focused incivility, understood as speech that targets people, while comparing two different classification schemes across three Transformer-based models: a *pure-comment classification* method vis-á-vis a *joint sentence-comment classification* method. The former trains a supervised objective on comment labels; the latter trains a joint supervised objective on both comment and sentence labels. The comparison is carried out on a relatively large, human-annotated, stratified dataset ($N = 7,941$) collected from a likely multicultural online setting, namely two war-related subreddits (r/IsraelPalestine and r/UkrainianConflict). The findings highlight small performance gains when training an objective supervised on both comment and sentence labels, and this gain is consistent across seeds and architectures (BERT, RoBERTa, and BERTweet).

Link to anonymized Github Repo: Link.

***Warning**: The study contains statements that may be offensive to some. These are fictive although paraphrased from most-common uncivil utterances in order to provide empirical examples.*

## 1 Introduction

Debates and discussions have long been considered an integral part of democratic participation, while the civility of such exchanges continues to receive considerable attention as much of the political talk has moved online. One important concept of anti-normative speech is incivility. Although contested, incivility is generally considered a communicative violation of social norms. Research shows that incivility and its variants are a cause of political polarization (Anderson et al., 2014; Neyazi et al., 2025; Suhay et al., 2018), political mistrust (Mutz and Reeves, 2005), negative emotions (Gervais, 2015), and even violent crime (Williams et al., 2019). Previously, scholars manually labeled datasets according to whether a given text or utterance was civil or uncivil. Recently, machine learning (ML) has improved large-scale detection of incivility by circumventing the time and monetary constraints associated with human annotations.

Yet, the promises of ML classifiers for the large-scale detection of incivility remain unclear in at least two respects. First, given the contested nature of the concept, most classifiers have thus far been trained on what Rossini (Rossini, 2022) considers tone-based indicators of incivility. This is inappropriate as tone-based indicators are, to a higher degree, dictated by the cultural context. Thus, focusing instead on content-based indicators of incivility allows researchers to home in on multicultural contexts, while focusing primarily on speech that has non-negligible effects on democracy.

Second, classifiers of incivility are trained on comments as the unit of analysis. While studies have achieved relatively high accuracy using this unit of analysis, it remains unknown how classifiers perform when using more granular units as auxiliary signals. This is problematic as anti-normative speech typically presents itself at more granular levels, such as sentences (de Gibert et al., 2018). Therefore, the paper asks the following question: Does joint sentence *and* comment supervision improve comment-level detection of content-based incivility across architectures?

To this end, we compare two classification schemes, using Transfomer-based models: comment-level classification, and comment-level classification using sentence labels as auxiliary signals. The data used to fine-tune these classifiers is a stratified sample ($N = 7,941$) of Reddit comments, collected from two war-related subreddits (r/IsraelPalestine and r/UkrainianConflict). The findings suggest that using sentence labels as auxiliary signals enhances commment-level classifica-

tion compared to pure-comment supervision. Although the increase in performance is marginal, it is relatively consistent across architectures and, for the most part, significant across seeds.

This work has at least three important contributions. Firstly, the paper goes beyond document classification and uses a multitask scheme where gradients pass through both comment and sentence labels. Secondly, the paper presents one of the first attempts to understand the content of anti-normative speech more broadly, making the approach particularly suitable across cultural contexts. Thirdly, the findings show small but significant performance gains over multiple seeds and across different architectures.

## 2 Related Work

Incivility is a type of anti-normative speech that, like many social science concepts, is fraught with contestations. As (Muddiman, 2017) explains, there are generally two schools of thought. On the one hand, some scholars associate incivility with disrespecting norms of inter-personal politeness norms (Mutz, 2015), such as name-calling (Anderson et al., 2014; Coe et al., 2014), swearing (Coe et al., 2014), and using multiple exclamation marks or all-caps (Gervais, 2016). On the other hand, some scholars see incivility as the disrespect against public-level norms, such as disrespecting the equal worth of all members or would-be members of discussion through stereotypical language (Papacharissi, 2004; Rowe, 2015) or outright hate speech (Chen, 2017).

In a recent but seminal paper, Patricia Rossini (Rossini, 2022) criticizes conceptualizations of incivility in their over-reliance on impoliteness. She argues that incivility-as-impoliteness is but a focus on the tone rather than the content of communication. Content vis-á-vis tone, as Zizi Papacharissi (Papacharissi, 2004) likewise argues, is the more important attribute of speech because (1) definitions of incivility should be robust across time and space-dependent contexts and (2) only content-based incivility has a non-negligible effect on democracy. However, at scale, detection of incivility has largely focused on tone-based indicators of incivility, including vulgar language (Davidson et al., 2020; Daxenberger et al., 2018; Gao et al., 2024; Ziegele et al., 2018).

Early ML approaches to detect incivility and hate speech at scale primarily used support vector machines (SVM), random forest decision trees, and logistic regression (Burnap and Williams, 2014; Davidson et al., 2017; de Gibert et al., 2018; Stoll et al., 2020). Since the earlier efforts, machine classification has become somewhat more sophisticated with most scholarly approaches nowadays utilizing deep learning techniques, such as CNNs, RNNs, and Transformer-based models (Davidson et al., 2020; Risch et al., 2019; Saleh et al., 2021). Moreover, recent efforts to detect anti-normative speech have also opted for the use of Large Language Models (LLMs) with zero- or few-shot prompting (Das et al., 2024; Dev et al., 2025; Huang et al., 2023; Roy et al., 2023).

Nevertheless, problems remain. For instance, Bonetti et al., 2023 find that the increase in performance with pre-trained LLMs may not be as big as imagined compared to the more traditional approaches. A comprehensive review (Albladi et al., 2025) finds that prompting methods suffer from inconsistent efficiency in detecting anti-normative speech across languages. Stoll et al., 2025 also find that LLM classification of incivility is fraught with societal biases.

Moreover, it remains unclear whether whole documents provide the best and only unit of analysis in comment-level classification of anti-normative speech. Most studies, by far, detect incivility or hate speech at the comment level. However, acknowledging the granularity of anti-normative speech alongside the necessity to make ML explainable, recent studies detect spans, i.e., spans of text that support a particular labeling decision, such as hate and toxic speech. Sarker et al., 2023, Paraschiv et al., 2024, and Pavlopoulos et al., 2022 find that span detection can be done with relatively high accuracy, while Williams et al., 2019 train hate speech classifiers alongside such spans. A handful of studies, e.g., Corso et al., 2024 and de Gibert et al., 2018, detect anti-normative speech at the sentence-level. As de Gibert et al., 2018 note, sentence and span classification rather than whole-document classification might be appropriate given that they represent the minimum unit of analysis where anti-normative speech occurs and that they avoid the noise of surrounding "civil" sentences.

Currently, no study has compared performance on comment-level classification between regimes that train on different levels of granularity. We agree with Rossini and Papacharissi that content matters more than tone and define incivility as the extent to which speech targets individuals or

groups. This includes (1) ad-hominem attacks, such as name-calling and mocking remarks about someone's communication, and (2) stereotypical language and outright hate speech. Such utterances may occur at any point in a comment for it to be classified as uncivil, with the minimum granularity of incivility signals being at the sentence-level. Content-based incivility requires the minimum unit of analysis to be at a higher dimension than tone-based incivility, which may be as low as individual tokens (e.g., "fuck"), since the focus on targets requires more of the semantic context.

We formalize the problem of comment classification as follows: Let the $j$-th comment be $D_j = (s_{j,1}, ..., s_{j,i}, ..., s_{j,n_j})$ with $n_j$ sentences. Each sentence $s_{j,i}$ has a binary label $y_{j,i} \in \{0, 1\}$ (e.g., 0='civil' and 1='uncivil'). Under the "at least one" rule, i.e., a given comment is a positive case if any of its sentences is a positive case:

$$y_j = \max_{i,...,n_j} y_{j,i} \tag{1}$$

such that:

$$y_j = 1 \Leftrightarrow \sum_{i=1}^{n_j} y_{j,i} \geq 1 \quad and$$
$$y_j = 0 \Leftrightarrow \sum_{i=1}^{n_j} y_{j,i} = 0$$

However, since sentence labels also depend on their local context, $y_{j,i}$ is a function of all sentences:

$$y_{j,i} = F(S_j, i) \tag{2}$$

where $S_j = (s_{j,1}, ..., s_{j,n_j})$. Then we can express $y_j$ as follows:

$$y_j = G(F(S_j, 1), ..., F(S_j, n_j)) = \max_{i=1,...,n_j} F(S_j, i) \tag{3}$$

such that:

$$y_j = 1 \Leftrightarrow \sum_{i=1}^{n_j} F(S_j, i) \geq 1 \quad and$$
$$y_j = 0 \Leftrightarrow \sum_{i=1}^{n_j} F(S_j, i) = 0$$

Suppose that we fine-tune a classifier for labeling comments, so that gradients flow through all tokens of a document, $D_j$. Suppose further that we have many long-form comments, some where all sentences are civil, some where all sentences are uncivil, and some where all but one sentence are civil. Using comment-level classification may lead to a civil classification in the latter case, although the 'ground truth,' as established by (1), is uncivil. The summary signal in this case is built from just one uncivil cue and many civil cues, which might average out the effect of the uncivil sentence on the comment's final classification. Thus, we risk that uncivil signals are crowded out by their civil counterparts, leading to false negatives (henceforth "dilution").

Alternatively, suppose then that we first fine-tune a sentence-level classifier for labeling sentences, where gradients only flow through the tokens of the sentence. After using the fine-tuned model to label sentences, we aggregate the labels at the comment level and decide based on (1). The problem here is that $y_{j,i}$ is now agnostic with respect to the local context:

$$y_{j,i} = F(i) \tag{4}$$

Once we aggregate sentence labels to obtain a comment label, the following form applies:

$$y_j = G(F(1), ..., F(n_j)) = \max_{i=1,...,n_j} F(i) \tag{5}$$

such that:

$$y_j = 1 \Leftrightarrow \sum_{i=1}^{n_j} F(i) \geq 1 \quad and$$
$$y_j = 0 \Leftrightarrow \sum_{i=1}^{n_j} F(i) = 0$$

Therefore, the final comment label is a function of the sum of its "parts," i.e., individual sentences, without considering the more holistic interplay between these sentences. As previous research suggests, the local context matters in the ground truth labels of anti-normative speech (Becker and Troschke, 2023; Pavlopoulos et al., 2020). Using (5) as a classification scheme might then lead to false negatives, in the case that the context dictates incivility, or false positives, in the case that the context dictates civility. Thus, dilution has been avoided at the expense of the local context.

To avoid dilution and at the same time preserve context, a hybrid strategy may balance the shortcomings. That is, instead of predicting sentence incivility in isolation or only learning from full-document labels, the classifier receives additional learning signals from sentence annotations while maintaining holistic document context.

| Sample | Krippendorph's $\alpha$ | Percentage Agreement |
|---|---|---|
| **Comment-level** | 0.8464 | 92.47% |
| **Sentence-level** | 0.8274 | 91.44% |

Table 1: Agreement between primary and secondary coders on incivility.

## 3 Data

We collected the data continuously from December 2024 to July 2025 from two war-related subreddits (r/IsraelPalestine and r/UkrainianConflict) using Python's PRAW library (Reddit's official API). These subreddits were chosen for further analysis due to Reddit's structural affordances, making them a prime breeding ground for political discussions but also incivility. Reddit is quasi-anonymous, which makes its selection partially ethical and partially due to analytical relevance. For decades, if not centuries, scholars have proved a link between the lack of immediate identifiability and anti-normative behavior, both in online as well as offline setting (Rowe, 2015; Solomon et al., 1978; Zimbardo, 1969). Finally, the war-related subreddits are particularly relevant for the given study objective as they encourage discussion of a non-domestic issue, likely inviting discussants from multiple cultural contexts and thereby problematizing the focus on tone rather than content when detecting incivility.

For each subreddit, an automated scraping agent retrieved the latest 200 posts and all corresponding comments twice daily, producing an initial dataset of 833,593 posts and comments (henceforth "comments" for simplicity). We opted for a continuous collection method, given the need to prevent high levels of eventual self-censorship and moderation in the case of uncivil speech. After removing known bots[1] as well as deleted or moderated content[2], the final dataset comprised 809,275 comments.

The cleaned dataset was then analyzed using PerspectiveAPI Toxicity Scores to construct a balanced sample, ensuring that the impending classifiers train on a larger range of uncivil instances. PerspectiveAPI defines toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" (Google, n.d.), with scores ranging from 0 (very unlikely to be toxic) to 1 (very likely to be toxic). A balanced sample of 8,000 comments was randomly drawn from the dataset, with 20% of cases from each of the following toxicity bands: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1. After removing links, embedded comments, and empty-text comments, the final sample contained 7,941 comments.

The balanced sample was annotated by a primary coder using a detailed codebook. To assess intercoder reliability, four PhD students independently coded 600 comments (approximately 8% of the balanced sample). See Table 1 for values of inter-coder reliability.

Both comments and posts in the balanced sample, as well as their individual sentences, were coded according to whether they were civil or uncivil. In total, the 7,941 comments contained 24,031 sentences.

Coding incivility is no straightforward task. The current paper adopts the idea put forward by Papacharissi, 2004 and later Rossini, 2022 that the focus of anti-normative speech should be that which has non-negligible effects on democracy. Therefore, the focus here is not on the tone, such as profanity, but rather on the content of speech, and incivility is defined as speech that negatively targets people.

Coders were thus asked to code comments and sentences as uncivil if it negatively targeted an individual or a group of people. As such, coders were instructed to label profane submissions as uncivil only if the profanity was other-directed, i.e., targeting either individuals or groups. Figure 1 presents some common examples of civility with group and individual targets.

Despite clear rules on the target, classification always carries an element of subjectivity (Chen et al., 2019) while speech at times is ambiguous. When context was missing or the incivility was deemed subtle, coders were asked to code positive cases of incivility if, beyond a reasonable doubt, the comment, or sentence negatively targets an individual or group of people. Moreover, sarcasm was coded as uncivil if its primary purpose was to belittle an individual or if it reproduced hateful or negative stereotypes.

[1] E.g., u/AutoModerator.

[2] 1.43% of the submissions ($N = 11,619$) were inaccessible due to deletion or moderation.

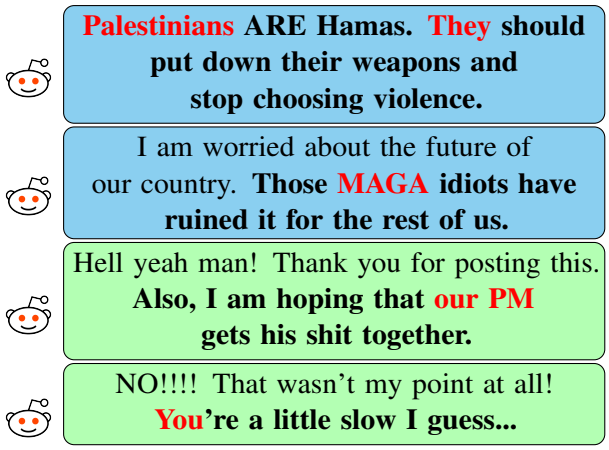


Figure 1: Examples of uncivil comments, where the red font color denotes the target and the bold text denotes which sentences of a given comment is uncivil. Targets are either groups of people as in the purple shaded boxes or individuals as in the green shaded boxes. Note: The examples are fictive but created from most-common uncivil comments in the corpus.

| | |
|---|---|
| **Avg. # of sentences** | 4.04 |
| **Toxicity score** $<$ **0.2** | 45.05% |
| **0.2** $\leq$ **Toxicity score** $<$ **0.4** | 27.76% |
| **0.4** $\leq$ **Toxicity score** $<$ **0.6** | 15.83% |
| **0.6** $\leq$ **Toxicity score** $<$ **0.8** | 8.50% |
| **0.8** $\leq$ **Toxicity score** | 2.93% |

Table 2: Civil sample, $N = 3,436$.

| | |
|---|---|
| **Avg. # of sentences** | 3.08 |
| **Avg. # of uncivil sentences** | 1.46 |
| **# without uncivil sentences** | 105 |
| **Toxicity score** $<$ **0.2** | 2.71% |
| **0.2** $\leq$ **Toxicity score** $<$ **0.4** | 12.50% |
| **0.4** $\leq$ **Toxicity score** $<$ **0.6** | 23.09% |
| **0.6** $\leq$ **Toxicity score** $<$ **0.8** | 28.81% |
| **0.8** $\leq$ **Toxicity score** | 32.90% |

Table 3: Univil sample, $N = 4,505$.

Descriptive statistics for the stratified sample are presented in Table 2 and 3. First, civil comments are composed of slightly more sentences than uncivil comments. Second, of the uncivil submissions, an average of 1.46 sentences are uncivil, while 105 submissions contain no uncivil sentences, i.e., while the overall comment is uncivil, no sentences are uncivil regardless of the local context. Third, there seems to be a strong but non-perfect relationship between toxicity and incivility, legitimizing the choice of training the classifiers on a stratified sample. Uncivil submissions are more likely to be very toxic, while only a few civil submissions (7.28%) fall in the very toxic range. This is likely because PerspectiveAPI's toxicity scores include tone-based indicators of anti-normative speech, such as profanity.

## 4 Experiments

In the following, we compare two methods of training classifiers for comment-level detection of incivility. We do this across three models: BERT, RoBERTa, and BERTweet. The first method, the *pure-comment method*, calculates loss only as a function of whether the model correctly labels the comments. Essentially, the model takes the comment as the only important unit of analysis for training. The second method, the *joint sentence-comment method*, runs two separate forward passes in each training step: one batch of comments for the comment loss, and one batch of independent sentences for the sentence loss. Here, the joint objective is the following:

$$\mathcal{L} = \mathcal{L}_{\text{comment}} + \alpha\, \mathcal{L}_{\text{sentence}}, \tag{6}$$

When $\alpha = 0$, the sentence supervision is effectively lost, and when $\alpha = 1$, the model updates loss based on sentence and comment supervision equally and jointly. Thus, the Transformer encoder is equivalent between the two classification schemes, while the difference lies in that the joint objective where both losses update the shared encoder parameters.

In the impending analysis, the pure-comment classification ($\alpha = 0$) is compared to the joint sentence-comment classification ($\alpha = 1$) across three transformer-based models. BERT represents the original Transformer encoder model and was trained with masked language modeling on a large general-domain corpus of English-language text (Devlin et al., 2019). RoBERTa has the same core architecture as BERT, while it was developed to optimize the original BERT pre-training procedure, i.e., by training longer with larger batches and by using dynamic masking (Liu et al., 2019). Finally, BERTweet (Nguyen et al., 2020) also retains the original BERT architecture, but it was trained on a large corpus of Twitter data. This might make it particularly suitable for the Reddit dataset by being attentive to social media and short-form content.

Before fine-tuning, the data was partitioned into a train (70%,$N = 5,558$), validation (15%, $N = 1,191$), and held-out test set (15%, $N = 1,192$).

| Hyperparameters | BERT, RoBERTa, BERTweet |
|---|---|
| **Learning Rate** | $10^{-5}$ |
| **Max Seq. Length** | 512 |
| **Batch Size** | 16 |
| **Epochs** | 30 |
| **Warm-up Proportion** | 0.10 |
| **Label Smoothing** | 0.15 |
| **Gradient Clipping** | 1.0 |
| **Seeds** | 20 runs with seeds $42 - 61$ incl. |

Table 4: Hyperparameters that generally yield best performance across models.

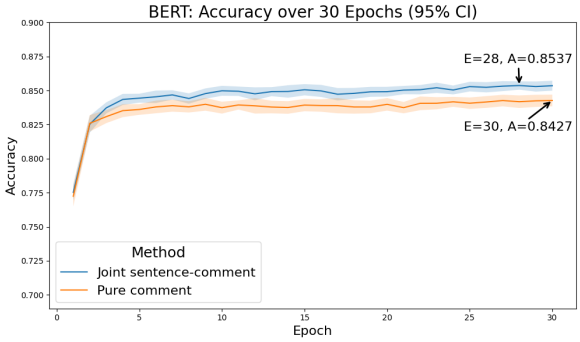


Figure 2: BERT accuracy on validation data. Shaded area represents 95% confidence intervals over 20 seeds.

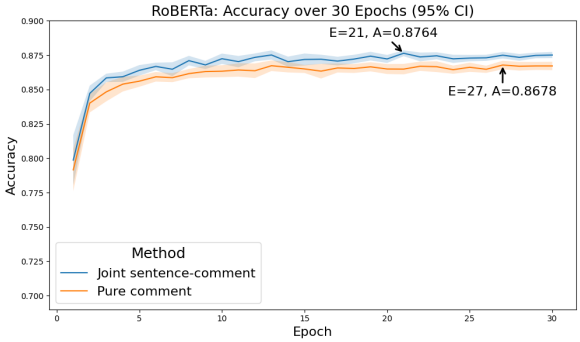


Figure 3: RoBERTa accuracy on validation data. Shaded area represents 95% confidence intervals over 20 seeds.

The text was then tokenized using the model-specific fast tokenizer, with padding to a maximum sequence length of 512 tokens and truncation enabled. We optimized with AdamW (LR=$10^{-5}$), and used a linear warmup-and-decay learning-rate schedule with a warmup proportion of 0.10 of the total number of training steps. Losses were computed with class-weighted cross entropy and label smoothing of 0.15. The gradients were clipped to max norm 1.0, and mixed-precision training was enabled on GPUs. Moreover, the performance of all model-method pairs were averaged over 20 seeds. Averaging over seeds avoids noise that arises randomly from dataset partitioning. Seed specifications and other hyperparameters are presented in Table 4.

## 5 Results

Figure 2, 3, and 4 depict the average-over-seeds performance on the validation set over 30 epochs. In all three figures, it is evident that the joint sentence-comment method performs consistently better than the pure-comment method. Although the difference is marginal, it is also significant at most epochs, provided by the lack of confidence interval overlaps.

Furthermore, Table 5 depicts the average-over-seeds performance of each model-method pair on a held-out test set, using the best-performing epoch. For all models, the joint sentence-comment classification performs better than the pure-comment classification. Though marginal, for both RoBERTa and BERTweet, this difference in accuracy is significant at the 95% level.

The best performing model-method pair is the BERTweet joint sentence-comment, which achieves an accuracy of 0.8831. This is relatively high compared to previous studies detecting anti-normative speech. Additionally, this is more than a 1 percentage-point increase compared to the BERTWeet pure-comment supervision. Besides the higher accuracy of the overall model-method pair, it also performs significantly better in terms of classifying both civil and uncivil comments compared to pure-comment supervision, as shown by the f1-scores for each category.

## 6 Conclusion

With advancements in ML methods, many scholars have turned to ML approaches in the detection of anti-normative speech, such as hate speech and incivility. With these more advanced methods, detection has become highly sophisticated, with high levels of performance. However, most scholarly approaches detect anti-normative speech at the comment level, leaving open the possibility of diluting out uncivil sentences with surrounding civil sentences. As a result of this, some scholars have opted for the detection of more granular levels of anti-normative speech, such as spans or sentences. We argue against this approach. As formalized above, while the comment-level classification is dependent on the signals of individual sentences,

| Model | A | $P_0$ | $R_0$ | $F1_0$ | $P_1$ | $R_1$ | $F1_1$ |
|---|---|---|---|---|---|---|---|
| **BERT, Pure comment** | 0.8456 (0.0026) | 0.8291 (0.0036) | 0.8128 (0.0059) | 0.8198 (0.0033) | 0.8596 (0.0037) | 0.8707 (0.0036) | 0.8649 (0.0022) |
| **RoBERTa, Pure comment** | 0.8643 (0.0023) | 0.8573 (0.0029) | 0.8236 (0.0053) | 0.8399 (0.0030) | 0.8697 (0.0034) | 0.8953 (0.0026) | 0.8822 (0.0019) |
| **BERTweet, Pure comment** | 0.8711 (0.0020) | 0.8524 (0.0044) | 0.8494 (0.0059) | 0.8511 (0.0024) | 0.8860 (0.0037) | 0.8877 (0.0046) | 0.8866 (0.0018) |
| **BERT, Joint sentence-comment** | 0.8519 (0.0018) | 0.8348 (0.0033) | 0.8203 (0.0051) | 0.8272 (0.0024) | 0.8652 (0.0031) | 0.8759 (0.0034) | 0.8703 (0.0016) |
| **RoBERTa, Joint sentence-comment** | 0.8747 (0.0027) | 0.8642 (0.0032) | 0.8410 (0.0053) | 0.8522 (0.0034) | 0.8815 (0.0034) | 0.8991 (0.0027) | 0.8901 (0.0022) |
| **BERTweet, Joint sentence-comment** | **0.8831** (0.0027) | **0.8715** (0.0039) | **0.8568** (0.0046) | **0.8636** (0.0032) | **0.8926** (0.0030) | **0.9035** (0.0034) | **0.8976** (0.0024) |

Table 5: Results on a held-out test set for all model-method pairs. Note: Standard errors in brackets; # of seeds per model-method pair = 20.

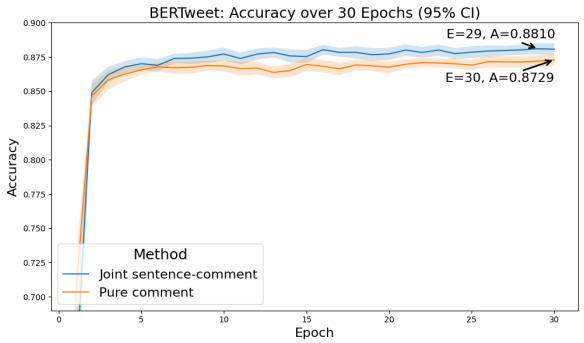


Figure 4: BERTweet accuracy on validation data. Shaded area represents 95% confidence intervals over 20 seeds.

it is not simply the sum of its "parts" since these sentences also depend on the local context.

Rather, this paper compares a pure-comment model to a joint sentence-comment model, the latter of which trains jointly on comments and sentence labels. The findings of this novel approach to detecting anti-normative speech suggest small but consistent performance gains when training a joint objective. This difference in performance is particularly pronounced and significant when using the BERTweet model. Finally, the paper provides relatively high performance of the detection of incivility, even when tone-related indicators of incivility are dropped, as well as a small but significant performance gain with the joint objective.

## 7 Limitations

**Architecture Generalizability**: This paper tested the performance gains in detecting incivility, a kind of anti-normative speech, when the objective is trained with attention to comment labels as well as sentence labels. While Transformer-based models represent a typical means of detecting anti-normative speech in the literature, the difference in performance between the two supervision regimes remains to be seen for other ML models, such as RNN, and CNNs.

**Data Generalizability**: Moreover, the findings are based on two war-related subreddits. This dataset was chosen given ethical considerations as well as its relevance to the content-focus in defining incivility. However, the combination of war-related discussion and the Reddit environment may provide a unique semantic and uncivil context that is not generalizable to other online political discussions.

**Choice of Alpha**: We compare training an objective on comment labels only and training an objective jointly on comment and sentence labels. Currently, the joint objective is strictly joint, giving equal preference to comment and sentence loss. We are therefore agnostic about the performance gains associated with varying $\alpha$ and thereby the importance that the models give to comment vis-á-vis sentence labels.

## 8 Ethical Statement

With the PRAW library, we used data from Reddit's official API to carry out the above analysis, which ensured compliance with Reddit's Terms of Service. The continuous scraping process also respected the API limits of 100 queries per minute. Moreover, the data was analysed without any identifiable information, such as usernames or IP addresses. Finally, we avoid the provision of any examples, with those provided in Figure 1 being purely fictive, although paraphrased from the most-common uncivil comments.

While the anonymized data will be made available to the community upon publication, we have

decided not to publish the LLMs fine-tuned with said data. This is first and foremost a legal consideration, given that Reddit prohibits the publication of LMMs fine-tuned with Reddit data.

The paper was approved by the university's ethical committee of the corresponding author.

## References


Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

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.

Matthias J. Becker and Hagen Troschke. 2023. Decoding implicit hate speech: The example of antisemitism. In Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe, editors, *Challenges and perspectives of hate speech research*, volume 12 of *Digital Communication Research*, pages 335–352. Leibniz-Institut für Sozialwissenschaften.

Andrea Bonetti, Marcelino Martínez-Sober, Julio C. Torres, Jose M. Vega, Sebastien Pellerin, and Joan Vila-Francés. 2023. Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*, 13(10).

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making. In *Political Science, Computer Science, Sociology*, pages 1–18.

Gina Masullo Chen. 2017. *Introduction: Incivility in Today's World*, chapter 1. Palgrave Macmillan.

Gina Masullo Chen, Ashley Muddiman, Tamar Wilner, Eli Pariser, and Natalie Jomini Stroud. 2019. We should not get rid of incivility online. *Social Media + Society*, 5(3):2056305119862641.

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.

Francesco Corso, Giuseppe Russo, and Francesco Pierri. 2024. A longitudinal study of italian and french reddit conversations around the russian invasion of ukraine. In *ACM Web Science Conference*, Websci '24, page 22–30. ACM.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2024. Evaluating ChatGPT against functionality tests for hate speech detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6370–6380, Torino, Italia. ELRA and ICCL.

Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

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.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20. ACL.

Tanni Dev, Sayma Sultana, and Amiangshu Bosu. 2025. Beyond binary moderation: Identifying fine-grained sexist and misogynistic behavior on github with large language models. *Preprint*, arXiv:2507.20358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Yujia Gao, Wenna Qin, Aniruddha Murali, Christopher Eckart, Xuhui Zhou, Jacob Daniel Beel, Yi-Chia Wang, and Diyi Yang. 2024. A crisis of civility? modeling incivility and its effects in political discourse online. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):408–421.

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

Bryan T. Gervais. 2016. More than mimicry? the role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research*, 29(3):384–405.

Google. n.d. How it works: Using machine learning to reduce toxicity online.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM*

*Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ashley Muddiman. 2017. Personal and public levels of political incivility. *International Journal of Communication*, 11:21.

Diana C. Mutz. 2015. *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton University Press.

Diana C. Mutz and Byron Reeves. 2005. The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, 99(1):1–15.

Taberez Ahmed Neyazi, Ozan Kuru, and Subhayan Mukerjee. 2025. The influence of sexism and incivility in whatsapp political discussions on affective polarization: Evidence from a 2022 multi-party election in india. *The International Journal of Press/Politics*, 30(3):797–818.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *Preprint*, arXiv:2005.10200.

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

Andrei Paraschiv, Teodora Andreea Ion, and Mihai Dascalu. 2024. Offensive text span detection in romanian comments using large language models. *Information*, 15(1).

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *Conference on Natural Language Processing*.

Patrícia Rossini. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425.

Ian Rowe. 2015. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2):121–138.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model. *Preprint*, arXiv:2111.01515.

Jaydeb Sarker, Sayma Sultana, Steven R Wilson, and Amiangshu Bosu. 2023. Toxispanse: An explainable toxicity detection in code review comments. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12. IEEE.

Linda Zener Solomon, William Neigher, and Henry Solomon. 1978. Anonymity and anti-normative behavior. *The Journal of Social Psychology*, 105(1):157–158.

Anke Stoll, Jingyuan Yu, Aliya Andrich, and Emese Domahidi. 2025. Classification bias of llms in detecting incivility towards female and male politicians in german social media discourse. *Communication Methods and Measures*, 19(4):350–368.

Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. Detecting impoliteness and incivility in online discussions: Classification approaches for german user comments. *Computational Communication Research*, 2(1):109–134.

Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. 2018. The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, 23(1):95–115.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Marc Ziegele, Johannes Daxenberger, Oliver Quiring, and Iryna Gurevych. 2018. Developing automated measures to predict incivility in public online discussions on the facebook sites of established news media. In *68th Annual Conference of the International Communication Association (ICA)*. ICA.

Phillip Zimbardo. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. *Nebraska Symposium on Motivation*, 17:237–307.