Ruddit: Norms of Offensiveness for English Reddit Comments

Rishav Hada[1,*], **Sohi Sudhir**[1,*], **Pushkar Mishra**[2], **Helen Yannakoudakis**[3], **Saif M. Mohammad**[4], **Ekaterina Shutova**[1]
[1]ILLC, University of Amsterdam
[2]Facebook AI, London
[3]Dept. of Informatics, King's College London
[4]National Research Council Canada
rishavhada@gmail.com, sohigre@gmail.com, pushkarmishra@fb.com,
helen.yannakoudakis@kcl.ac.uk, saif.mohammad@nrc-cnrc.gc.ca, e.shutova@uva.nl

Abstract

Warning: *This paper contains comments that may be offensive or upsetting.*

On social media platforms, hateful and offensive language negatively impact the mental well-being of users and the participation of people from diverse backgrounds. Automatic methods to detect offensive language have largely relied on datasets with categorical labels. However, comments can vary in their degree of offensiveness. We create the first dataset of English language Reddit comments that has *fine-grained, real-valued scores* between -1 (maximally supportive) and 1 (maximally offensive). The dataset was annotated using *Best–Worst Scaling*, a form of comparative annotation that has been shown to alleviate known biases of using rating scales. We show that the method produces highly reliable offensiveness scores. Finally, we evaluate the ability of widely-used neural models to predict offensiveness scores on this new dataset.

1 Introduction

Social media platforms serve as a medium for exchange of ideas on a range of topics, from the personal to the political. This exchange can, however, be disrupted by offensive or hateful language. Such language is pervasive online (Statista, 2020b), and exposure to it may have numerous negative consequences for the victim's mental health (Munro, 2011). Automated offensive language detection has thus been gaining interest in the NLP community, as a promising direction to better understand the nature and spread of such content.

There are several challenges in the automatic detection of offensive language (Wiedemann et al., 2018). The NLP community has adopted various definitions for offensive language, classifying it into specific categories. For example, Waseem and Hovy (2016) classified comments as *racist, sexist, neither*; Davidson et al. (2017) as *hate-speech, offensive but not hate-speech, neither offensive nor hate-speech* and Founta et al. (2018) as *abusive, hateful, normal, spam*. Schmidt and Wiegand (2017); Fortuna and Nunes (2018); Mishra et al. (2019); Kiritchenko and Nejadgholi (2020) summarize the different definitions. However, these categories have significant overlaps with each other, creating ill-defined boundaries, thus introducing ambiguity and annotation inconsistency (Founta et al., 2018). A further challenge is that after encountering several highly offensive comments, an annotator might find subsequent moderately offensive comments to not be offensive (*de-sensitization*) (Kurrek et al., 2020; Soral et al., 2018).

At the same time, existing approaches do not take into account that comments can be offensive to a different degree. Knowing the degree of offensiveness of a comment has practical implications, when taking action against inappropriate behaviour online, as it allows for a more fine-grained analysis and prioritization in moderation.

The representation of the offensive class in a dataset is often boosted using different strategies. The most common strategy used is key-word based sampling. This results in datasets that are rich in explicit offensive language (language that is unambiguous in its potential to be offensive, such as those using slurs or swear words (Waseem et al., 2017)) but lack cases of implicit offensive language (language with its true offensive nature obscured due to lack of unambiguous swear words, usage of sarcasm or offensive analogies, and others (Waseem et al., 2017; Wiegand et al., 2021)) (Waseem, 2016; Wiegand et al., 2019). key-word based sampling often results in spurious correlations (e.g., sports-related expressions such as *announcer* and *sport* occur very frequently in offensive tweets). Lastly, existing datasets consider of-

[*]Both authors contributed equally.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2700–2717
August 1–6, 2021.

fensive comments in isolation from the wider conversation of which they are a part. Offensive language is, however, inherently a social phenomenon and its analysis has much to gain from taking the conversational context into account (Gao and Huang, 2017).

In this paper, we present the first dataset of 6000 English language Reddit comments that has *fine-grained, real-valued scores* between -1 (maximally supportive) and 1 (maximally offensive) – normative offensiveness ratings for the comments. For the first time, we use comparative annotations to detect offensive language. In its simplest form, comparative annotations involve giving the annotators two instances at a time, and asking which exhibits the property of interest to a greater extent. This alleviates several annotation biases present in standard rating scales, such as scale-region bias (Presser and Schuman, 1996; Asaadi et al., 2019), and improves annotation consistency (Kiritchenko and Mohammad, 2017). However, instead of needing to annotate N instances, one now needs to annotate N^2 instance pairs—which can be prohibitive. Thus, we annotate our dataset using an efficient form of comparative annotation called *Best–Worst Scaling (BWS)* (Louviere, 1991; Louviere et al., 2015; Kiritchenko and Mohammad, 2016, 2017).

By eliminating different offensiveness categories, treating offensiveness as a continuous dimension, and eliciting comparative judgments from the annotators (based on their understanding of what is offensive), we alleviate the issues regarding category definitions and arbitrary category boundaries discussed earlier. By obtaining real-valued offensiveness scores, different thresholds can be used in downstream applications to handle varying degrees of offensiveness appropriately. By framing the task as a comparative annotation task, we obtain consistent and reliable annotations. We also greatly mitigate issues of annotator de-sensitization as one will still be able to recognize if one comment is more offensive than another, even if they think both comments are not that offensive.

In contrast to existing resources, which provide annotations for individual comments, our dataset includes conversational context for each comment (i.e. the Reddit thread in which the comment occurred). We conduct quantitative and qualitative analyses of the dataset to obtain insights into how emotions, identity terms, swear words, are related to offensiveness. Finally, we benchmark several widely-used neural models in their ability to predict offensiveness scores on this new dataset.[1]

2 Related Work

2.1 Offensive Language Datasets

Surveys by Schmidt and Wiegand (2017); Fortuna and Nunes (2018); Mishra et al. (2019); Vidgen and Derczynski (2020) discuss various existing datasets and their compositions in detail. Waseem and Hovy (2016); Davidson et al. (2017); Founta et al. (2018) created datasets based on Twitter data. Due to prevalence of the non-offensive class in naturally-occurring data (Waseem, 2016; Founta et al., 2018), the authors devised techniques to boost the presence of the offensive class in the dataset. Waseem and Hovy (2016) used terms frequently occurring in offensive tweets, while Davidson et al. (2017) used a list of hate-related terms to extract offensive tweets from the Twitter search API. Park et al. (2018), Wiegand et al. (2019), and Davidson et al. (2019) show that the Waseem and Hovy (2016) dataset exhibits topic bias and author bias due to the employed sampling strategy. Founta et al. (2018) boosted the representation of offensive class in their dataset by analysing the sentiment of the tweets and checking for the presence of offensive terms. In our work, we employ a hybrid approach, selecting our data in three ways: specific topics, emotion-related key-words, and random sampling.

Past work has partitioned offensive comments into *explicitly offensive* (those that include profanity—swear words, taboo words, or hate terms) and *implicitly offensive* (those that do not include profanity) (Waseem et al., 2017; Caselli et al., 2020a; Wiegand et al., 2021). Some other past work has defined explicitly and implicitly offensive instances a little differently: Sap et al. (2020) considered factors such as obviousness, intent to offend and biased implications, Breitfeller et al. (2019) considered factors such as the context and the person annotating the instance, and Razo and Kübler (2020) considered the kind of lexicon used. Regardless of the exact definition, implicit offensive language, due to a lack of lexical cues, is harder to classify not only for computational models, but also for humans. In our work, we consider implicitly offensive comments as those offensive comments that do not contain any swear words.

[1]Dataset and code available at: https://github.com/hadarishav/Ruddit.

Wulczyn et al. (2016, 2017) created three different datasets from Wikipedia Talk pages, focusing on aggression, personal attacks and toxicity. The comments were sampled at random from a large dump of English Wikipedia, and boosted by including comments from blocked users. For the personal attacks dataset, Wulczyn et al. (2016) used two different kinds of labels: ED (empirical distribution), OH (one hot). In case of ED, the comments were assigned real-valued scores between 0 and 1 representing the fraction of annotators who considered the comment a personal attack. While these labels were introduced to create a separation between the nature of comments with a score of 1.0 and those with a score of 0.6 (which would otherwise be classified as attacks), they are discrete. In our work, using the BWS comparative annotation setup, we assign fine-grained continuous scores to comments to denote their degree of offensiveness.

2.2 Best–Worst Scaling (BWS)

BWS was proposed by Louviere (1991). Kiritchenko and Mohammad (2017) have experimentally shown that BWS produces more reliable fine-grained scores than the scores acquired utilizing rating scales. In the BWS annotation setup, the annotators are given an n-tuple (where n $>$ 1, and commonly $n = 4$), and asked which item is the best and which is the worst (best and worst correspond to the highest and the lowest with respect to a property of interest). Best–worst annotations are particularly efficient when using 4-tuples, as each annotation results in inequalities for 5 of the 6 item pairs. For example, a 4-tuple with items A, B, C, and D, where A is the best, and D is the worst, results in inequalities: A$>$B, A$>$C, A$>$D, B$>$D, and C$>$D. Real-valued scores of associations are calculated between the items and the property of interest from the best–worst annotations for a set of 4-tuples (Orme, 2009; Flynn and Marley, 2014). The scores can be used to rank items by the degree of association with the property of interest. Within the NLP community, BWS has thus far been used only for creating datasets for relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), word–sentiment intensity (Kiritchenko et al., 2014), phrase sentiment composition (Kiritchenko and Mohammad, 2016), and tweet-emotion intensity (Mohammad and Bravo-Marquez, 2017; Mohammad and Kiritchenko, 2018). Using BWS, we create the first dataset with degree of offensiveness scores for social media comments.

3 Data collection and sampling

We extracted Reddit data from the Pushshift repository (Baumgartner et al., 2020) using *Google BigQuery*. Reddit is a social news aggregation, web content rating, and discussion website. It contains forums called *subreddits* dedicated to specific topics. Users can make a *post* on the subreddit to start a discussion. Users can *comment* on existing posts or comments to participate in the discussion. As users can also reply to a comment, the entire discussion has a hierarchical structure called the comment *thread*. We divided the extracted comments into 3 categories based on their subreddit source:

1. **Topics (50%)**: Contains comments from topic-focused subreddits: *AskMen, AskReddit, TwoXChromosomes, vaxxhappened, worldnews, worldpolitics*. These subreddits were chosen to cover a diverse range of topics. *AskReddit, vaxxhappened, worldnews, worldpolitics* discuss generic themes. *TwoXChromosomes* contains women's perspectives on various topics and *AskMen* contains men's perspectives.
2. **ChangeMyView (CMV) (25%)**: The CMV subreddit (with over a million users) has posts and comments on controversial topics.
3. **Random (25%)**: Contains comments from random subreddits.

We selected 808 posts from the subreddits based on criteria such as date, thread length, and post length. (Further details in the Appendix A.1.) We took the first 25 and the last 25 comments per post (skipping comments that had [DELETED] or [REMOVED] as comment body). The first responses are likely to be most relevant to the post. The final comments indicate how the discussion ended. We sampled 6000 comments from this set for annotation.

The goal of the sampling was to increase the proportion of offensive and emotional comments. Emotions are highly representative of one's mental state, which in turn are associated with their behaviour (Poria et al., 2019). For example, Jay and Janschewitz (2008) show that people tend to swear when they are angry, frustrated or anxious.

Studies have shown that the primary dimensions of emotion are valence, arousal, and dominance (VAD) (Osgood et al., 1957; Russell, 1980, 2003).

Valence is the positive – negative or pleasure–displeasure dimension. ***Arousal*** is the excited–calm or active–passive dimension. ***Dominance*** is powerful–weak or 'have full control'–'have no control' dimension (Mohammad, 2018). To boost the representation of offensive and emotional comments in our dataset, we up-sampled comments that included low-valence (highly negative) words and those that included high-arousal words (as per the NRC VAD lexicon (Mohammad, 2018)).[2] The manually constructed NRC VAD lexicon includes 20,000 English words, each with a real-valued score between 0 and 1 in the V, A, D dimensions.

In order to do this upsampling, we first defined the valence score of each comment as the average valence score of the negative words within the comment (A negative word is defined as a word with a valence score ≤ 0.25 in the VAD lexicon.) Similarly, we defined the arousal score for a comment as the average arousal score of high-arousal words in each comment. (A high-arousal word is defined as a word with an arousal score ≥ 0.75.)

We selected comments from the comment pool such that 50% were from the *Topics* category, 25% from the *CMV* category, and 25% from the *Random* category. Within each category, 33% of the comments were those that had the lowest valence scores, 33% of the comments were those that had the highest arousal scores, and the remaining were chosen at random.

4 Annotation

The perception of 'offensiveness' of a comment can vary from person to person. Therefore, we used crowdsourcing to annotate our data. Crowdsourcing helps us get an aggregation of varied perspectives rather than expert opinions which can leave out offensiveness in a comment that lies outside the 'typical' offensiveness norms (Blackwell et al., 2017). We carried out all the annotation tasks on Amazon Mechanical Turk (AMT). Due to the strong language, an adult content warning was issued for the task. Reddit is most popular in the US, which accounts for 50% of its desktop traffic (Statista, 2020a). Therefore, we restricted annotators to those residing in the US. To maintain the quality of annotations, only annotators with high approval rate were allowed to participate.

[2]In some initial pilot experiments, we found this approach of sampling low valence and high arousal comments to result in a greater number of offensive comments.

4.1 Annotation with Best–Worst Scaling

We followed the procedure described in Kiritchenko and Mohammad (2016) to obtain BWS annotations. Annotators were presented with 4 comments (4-tuple) at a time and asked to select the comment that is most offensive (least supportive) and the comment that is least offensive (most supportive). We randomly generated 2N distinct 4-tuples (where N is the number of comments in the dataset), such that each comment was seen in eight different 4-tuples and no two 4-tuples had more than 2 items in common. We used the script provided by Kiritchenko and Mohammad (2016) to obtain the 4-tuples to be annotated.[3]

Kiritchenko and Mohammad (2016) show that in a word-level sentiment task, using just three annotations per 4-tuple produces highly reliable results. However, since we work with long comments and a relatively more difficult task, we got each tuple annotated by 6 annotators. Since each comment is seen in 8 different 4-tuples, we obtain 8 X 6 = 48 judgements per comment.

4.2 Annotation Task and Process

In our instructions to the annotators, we defined offensive language as comments that include but are not limited to [being hurtful (with or without the usage of abusive words)/ being intentionally harmful/ treating someone improperly/ harming the 'self-concept' of another person/ aggressive outbursts/ name calling/ showing anger and hostility/ bullying/ hurtful sarcasm]. We also encouraged the annotators to follow their instincts. By framing the task in terms of comparisons and providing a broad definition of offensiveness, we avoided introducing artificial categories and elicit responses guided by their intuition of the language.

Detailed annotation instructions are made publicly available (Figure 5 in Appendix A.2).[4] A sample questionnaire is shown in Figure 6 in Appendix A.2. For quality control purposes, we manually annotated around 5% of the data ourselves beforehand. We will refer to these instances as *gold questions*. The gold questions were interspersed with the other questions. If a worker's accuracy on the gold questions fell below 70%, they were refused further annotation and all of their annotations were discarded. The discarded annotations were

[3]http://saifmohammad.com/WebPages/BestWorst.html

[4]AMT task interface with instructions: https://hadarishav.github.io/Ruddit/

# Comments	# Annotations per Tuple	# Annotations	# Annotators	SHR Pearson	SHR Spearman
6000	6	95,255	725	0.8818 ± 0.0023	0.8612 ± 0.0029

Table 1: Ruddit annotation statistics and split-half reliability (SHR) scores.

published again for re-annotation. We received a total of 95,255 annotations by 725 crowd workers.

The BWS responses were converted to scores using a simple counting procedure (Orme, 2009; Flynn and Marley, 2014). For each item, the score is the proportion of times the item is chosen as the most offensive minus the proportion of times the item is chosen as the least offensive. We release the aggregated annotations as well as the individual annotations of Ruddit, to allow further work on examining and understanding the variability.[5]

4.3 Annotation Reliability

We cannot use standard inter-annotator agreement measures to ascertain the quality of comparative annotations. The disagreement that arises in tuples having two items that are close together in their degree of offensiveness is a useful signal for BWS (helping it give similar scores to the two items). The quality of annotations can be measured by measuring the reproducibility of the end result – if repeated manual annotations from multiple annotators can produce similar rankings and scores, then, one can be confident about the quality of annotations received. To assess this reproducibility, we computed average *split-half reliability* (SHR) values over 100 trials. SHR is a commonly used approach to determine consistency in psychological studies.

For computing SHR values, the annotations for each 4-tuple were randomly split in two halves. Using these two splits, two sets of rankings were determined. We then calculated the correlation values between these two sets. This procedure was repeated 100 times and the correlations were averaged. A high correlation value indicates that the annotations are of good quality. Table 1 shows the SHR for our annotations. SHR scores of over 0.8 indicate substantial reliability.

5 Data Analysis

In this section, we analyze various aspects of the data, including: the distribution of scores, the association with identity terms, the relationship with emotion dimensions, the relationship with data source, and the role of swear words.

[5]We provide the comment IDs and not the comment body, in accordance to the GDPR regulations. Comment body can be extracted using the Reddit API.

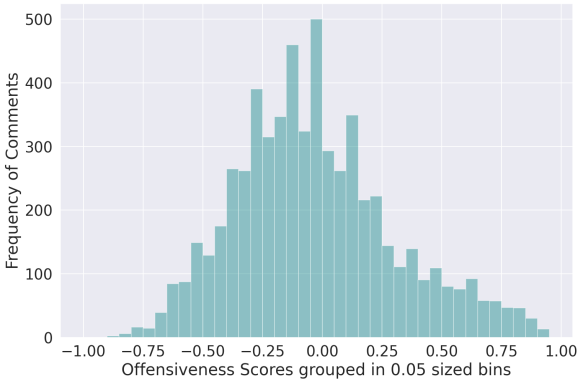

Figure 1: A histogram of frequency of comments–degree of offensiveness. Degree of offensiveness scores are grouped in bins of size 0.05.

Distribution of Scores Figure 1 shows a histogram of frequency of comments vs. degree of offensiveness, over 40 equi-spaced score bins of size 0.05. We observe a normal distribution.

To analyze the data, we placed the comments in 5 equi-spaced score bins of size 0.4 (bin 1: -1.0 to -0.6, bin 2: -0.6 to -0.2, and so on). Table 2 shows some comments from the dataset (more examples can be found in Appendix A.3 Table 6). We observed that bin 1 primarily contains supportive comments while bin 2 shows a transition from supportive to neutral comments. Bin 3 is dominated by neutral comments but as the score increases the comments become potentially offensive and bins 4 & 5 predominantly contain offensive comments. It is interesting to note that bin 4 contains some instances of implicit offensive language such as *'You look like a lesbian mechanic who has a shell collection'*. In their paper, Wiegand et al. (2021) explore the category of such "*implicity abusive comparisons*", in depth. More examples of implicitly offensive comments present in our dataset can be found in table 2 and table 6 (in Appendix A.3).

To explore whether specific bins capture specific topics or key-words, we calculated Pointwise Mutual Information (PMI) scores of all the unique words in the comments (excluding stop words) with

Bin	Comment	Score
1	Don't worry, she's going to be fine.	−0.75
	I see you too are a man of culture;)	−0.604
2	This is so sexy! Love it!	−0.562
	"I live with my ex, but it's totally cool, we're just friends"	−0.229
3	Not sure why Im being down voted? Why does the truth bother so many people?	−0.191
	I presented a hypothetical question to you. I did not even claim that you made that argument. Unfortunately that is not a straw man. So, care to answer that question again?	0.083
4	Don't forget Vaccines cause autism. And torture is awesome. We should murder the families of terrorists.	0.5
	What is your angle, Kim?? Is this some Hitler BS where you sign a peace treaty and then start WWIII? Or did you finally just grow a brain? Because neither sound particularly more likely than the other...	0.521
5	If you support trump kill yourself, painfully	0.604
	shut the fuck up bitch. It's Bernie or Bust nobody is voting for Biden, now get the fuck out of here you cunt	0.958

Table 2: Sample comments from Ruddit for each of the 5 score bins. Comment in bold is implicitly offensive.

Bin	Words
1	awesome, thanks, appreciate
2	songs, headphones, sweet, movie
3	gap, sacrifice, employee
4	muslim, fucked, gay, ass, raped
5	dick, fuck, asshole, ass, shut

Table 3: Top PMI scoring words for each of the 5 offensiveness-score bins. Degree of offensiveness scores are grouped in bins of size 0.4.

the five score bins. Table 3 shows the top scoring words for each bin. We observed that bins 1, 2, and 3 exhibit a strong association with supportive or neutral words, while bins 4 and 5 show a strong association with swear words and identity terms commonly found in offensive contexts.

Identity terms A common criticism of the existing offensive language datasets is that in those datasets, certain identity terms (particularly those referring to minority groups) occur mainly in texts that are offensive (Sap et al., 2019; Davidson et al., 2019; Wiegand et al., 2019; Park et al., 2018; Dixon et al., 2018). This leads to high association of targeted minority groups (such as Muslims, females, black people and others) with the offensive class(es). This bias, in turn, is captured by the computational models trained on such datasets. As mentioned earlier, in Ruddit, certain words such as *gay, trans, male, female, black, white* were found to exhibit a relatively higher association with the offensive bins than with the supportive bins. In order to probe the effect of this on the computational models, we created a variant of Ruddit by replacing all the identity terms (from the list given in Appendix A.4) in the comments with the [*group*] token and observed the effect on the models' performance. We refer to this variant of the dataset as the ***identity-agnostic dataset***. We analyse the models' performance in the next section.

Offensiveness vs. emotion As discussed earlier, our emotions impact the words we use in text. We examined this relationship quantitatively using Ruddit and the NRC VAD Lexicon (which has intensity scores along the valence, arousal, and dominance dimensions). For each comment in Ruddit, we calculated three scores that captured the intensities of the V, A, D words (the averages of the intensities of the V/A/D words in the comment), using the entire lexicon. We then determined the correlation between each of the three scores and the degree of offensiveness. Only comments containing at least 4 words from the VAD lexicon were considered for the score and correlation calculation. A total of 4831 comments qualified the criteria. See Table 4. From the table, we can observe that *valence* is weakly inversely correlated, *arousal* is weakly correlated, and *dominance* does not exhibit notable correlation with offensiveness. This behaviour can also be observed in Figure 2 that shows a plot of the average V, A, and D scores of comments in the five equi-spaced offensiveness-score bins. Note the clear trend that as we look at bins with more offensive comments, the average valence of the comments decreases and the average arousal increases.

Offensiveness vs. data source As mentioned earlier, comments in our dataset come from three different sources - Topics, CMV, and Random. Figure 3 shows the distribution of comments from each source over the score bins. We observed that comments from Topics have near equal representation on both sides of the scale, while for the other

Emotion	Pearson's r
Valence	−0.301
Arousal	0.256
Dominance	−0.086

Table 4: Pearson correlation values between the offensiveness scores and the emotion dimension scores.

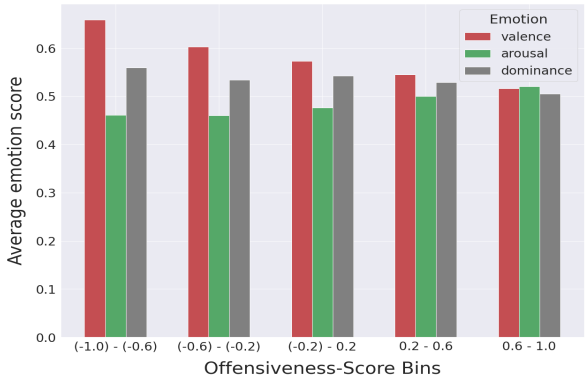

Figure 2: Average Valence, Arousal, and Dominance scores of comments in various offensiveness-score bins.

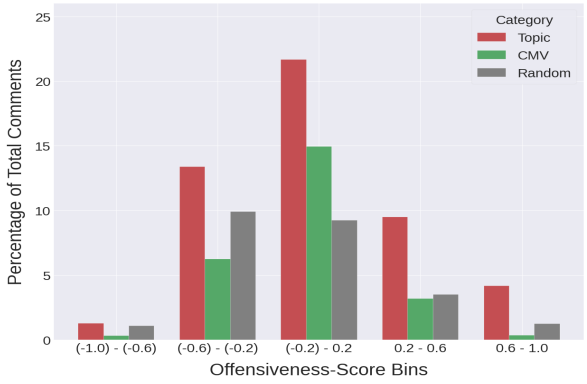

Figure 3: Distribution of comments from each comment category in each offensiveness-score bin.

two sources, comments are more prevalent in the supportive bins. The higher representation of comments from Topics than the other two sources in the offensive bins, is likely due to the fact that the Topics category includes subreddits such as *worldnews* and *worldpolitics*. Discussions on these subreddits covers controversial topics and lead to the usage of offensive language. We observed that *worldnews* and *worldpolitics* indeed have high representation in the offensive bins (Figure 8 in Appendix A.4).

Swear words We identified 868 comments in our dataset that contain at least one swear word from the cursing lexicon (Wang et al., 2014). Comments containing swear words can have a wide range of offensiveness scores. To visualize the distribution, we plot a histogram of the comments containing swear words vs. degree of offensiveness (see Figure 7 in Appendix A.4). The distribution is skewed towards the offensive end of the scale. An interesting observation is that some comments with low offensiveness scores contain phrases using swear words to express enthusiasm or to lay more emphasis, for example '*Hell yes*', '*sure as hell love it*', '*uncomfortable as shit*' and others. To study the impact of comments containing swear words on computational models, we created another variant of Ruddit in which we removed all the comments containing at least one swear word. We refer to this variant as the ***no-swearing dataset***. This dataset contains 5132 comments. We analyse the models' performance on this dataset in the next section.

Offensiveness in different score ranges It is possible that comments in the middle region of the scale may be more difficult for the computational models. Thus, we created a subset of Ruddit containing comments with scores from −0.5 to 0.5. We call this subset (of 5151 comments), the ***reduced-range dataset***. We discuss the models' performance on this dataset in the next section.

6 Computational Modeling

In this section, we present benchmark experiments on Ruddit and its variants by implementing some commonly used model architectures. The task of the models was to predict the offensiveness score of a given comment. We performed 5-fold cross-validation for each of the models.[6]

6.1 Models

Bidirectional LSTM We fed pre-trained 300 dimensional GloVe word embeddings (Pennington et al., 2014) to a 2-layered BiLSTM to obtain a sentence representation (using a concatenation of the last hidden state from the forward and backward direction). This sentence representation was then passed to a linear layer with a *tanh* activation to produce a score between −1 and 1. We used *Mean Squared Error* (MSE) loss as the objective function, Adam with 0.001 learning rate as the optimizer, hidden dimension of 256, batch size of 32, and a dropout of 0.5. The model was trained for 7 epochs.

[6]Since we have a linear regression task, we created folds using *sorted stratification* (Lowe, 2016) to ensure that the distribution of all the partitions is similar.

Dataset	HateBERT		BERT		BiLSTM	
	r	MSE	r	MSE	r	MSE
a. Ruddit	0.886 ± 0.003	0.025 ± 0.001	0.873 ± 0.005	0.027 ± 0.001	0.831 ± 0.005	0.035 ± 0.001
b. *Identity-agnostic*	0.883 ± 0.006	0.025 ± 0.001	0.869 ± 0.007	0.027 ± 0.001	0.824 ± 0.007	0.036 ± 0.001
c. *No-swearing*	0.808 ± 0.013	0.023 ± 0.001	0.783 ± 0.012	0.027 ± 0.001	0.704 ± 0.014	0.036 ± 0.002
d. *Reduced-range*	0.781 ± 0.014	0.022 ± 0.001	0.757 ± 0.011	0.025 ± 0.001	0.659 ± 0.008	0.033 ± 0.001

Table 5: Five-fold cross-validation results of the models on Ruddit and its variants. r = Pearson's R. Note: Scores for c. and d. are not directly comparable to scores for a. and b. as they involve different score ranges.

BERT We fine-tuned BERT$_{base}$ (Devlin et al., 2019). We added a regression head containing a linear layer to the pre-trained model. We used MSE loss as the objective function, batch size of 16, and learning rate of $2e-5$ (other hyperparameters same as (Devlin et al., 2019)). We used the *AdamW* optimizer with a linear learning rate scheduler with no warm up steps. The model was trained for 3 epochs. (More details in Appendix A.5.)

HateBERT HateBERT (Caselli et al., 2020b) is a version of BERT pretrained for abusive language detection in English. HateBERT was trained on RAL-E, a large dataset of English language Reddit comments from communities banned for being offensive or hateful. HateBERT has been shown to outperform the general purpose BERT model on the offensive language detection task when fine-tuned on popular datasets such as OffensEval 2019 (Zampieri et al., 2019), AbusEval (Caselli et al., 2020a), and HatEval (Basile et al., 2019).

We fine-tuned HateBERT on Ruddit and its variants. The experimental setup for this model is the same as that described for the BERT model.

6.2 Results and Analysis

We report Pearson correlation (r) and MSE, averaged over all folds. The performance of the models on Ruddit and its variants is shown in the Table 5. Note that the performance values on the *no-swearing* and the *reduced-range* datasets are not directly comparable to the performance values on the full Ruddit as their score range is different. We can see that on all the datasets, the HateBERT model performs the best, followed by the BERT model. Interestingly, the model performance (for all models) does not change substantially when trained on Ruddit or the *identity-agnostic* dataset. This indicates that the computational models are not learning to benefit from the association of certain identity terms with a specific range of scores on the offensiveness scale.[7]

The models show a performance drop on the *no-swearing* dataset, which suggests that swear words are useful indicators of offensiveness and that the comments containing them are easier to classify. Yet, the fact that the models still obtain performance of up to 0.8 (r) demonstrates that they necessitate and are able to learn other types of offensiveness features. It is also worth mentioning that even if they encounter swear words in a comment, the task is not simply to label the comment as offensive but to provide a suitable score.

Finally, the models obtained the performance of up to 0.78 (r) on the *reduced-range* dataset, which shows that even if the comments from the extreme ends of the offensiveness scale are removed, Ruddit still presents an interesting and feasible offensiveness scoring task.

Error Analysis Figure 4 shows the squared error values of the 3 models over the offensiveness score range in Ruddit. As expected, for all the models, the error in predictions is lower on both the extreme ends of the scale than in the middle region. Comments with very high or very low offensiveness scores are rich in obvious linguistic cues, making it easier for the computational models to predict scores. Most of the not-obvious, indirect implicitly offensive, and neutral comments should be present in the middle region of the offensiveness scale, making them more difficult for the models. It is interesting to observe that HateBERT, unlike the other two models, does not have high error values for samples within the score range 0.25–0.75. This indicates that HateBERT is efficient in dealing with offensive language that does not lie in the extreme offensive end. BiLSTM seems relatively less accurate for samples in the supportive range (-0.75 to -0.25). This could be attributed to the less complex model architecture and the usage of GloVe

[7] It should be noted that since the list of identity terms and the cursing lexicon we use is not exhaustive, our conclusions are only limited to the scope of the respective lists.

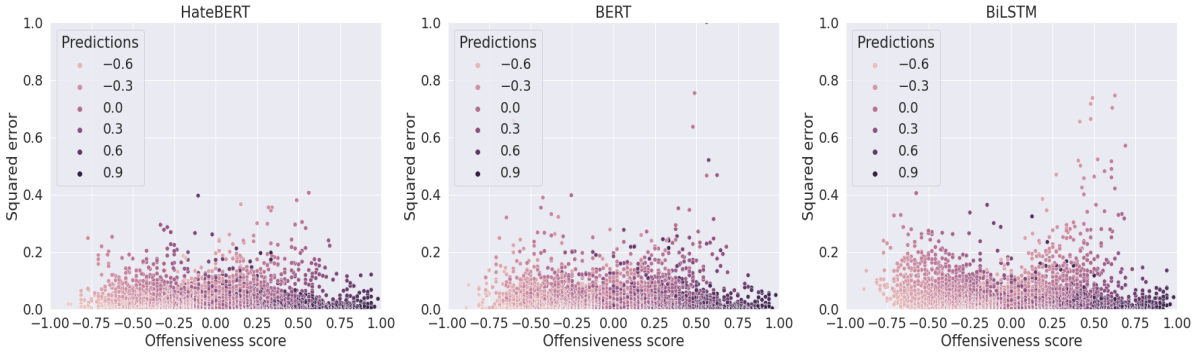

Figure 4: Squared error values for the 3 models' predictions over the offensiveness score range in Ruddit.

word embeddings.

7 Conclusion

We presented the first dataset of online comments annotated for their degree of offensiveness. We used a comparative annotation technique called Best–Worst Scaling, which addresses the limitations of traditional rating scales. We showed that the ratings obtained are highly reliable (SHR Pearson $r \approx 0.88$). We performed data analysis to gain insight into the relation of emotions, data sources, identity terms, and swear words with the offensiveness scores. We showed that valence is inversely correlated with offensiveness and arousal is directly correlated with offensiveness. Finally, we presented benchmark experiments to predict the offensiveness score of a comment, on our dataset. We found that computational models are not benefiting from the association of identity terms with specific range of scores on the offensiveness scale. In future work, it would be interesting to explore the use of conversational context in computational modeling of offensiveness, as well as studying the interaction between offensiveness and emotions in more depth. We make our dataset freely available to the research community.

Acknowledgements

This research was funded by the Facebook Online Safety Benchmark Research award for the project "A Benchmark and Evaluation Framework for Abusive Language Detection."

Ethical Considerations

We create Ruddit to study, understand and explore the nature of offensive language. Any such dataset might also be used to create automatic offensive language detection systems. While we realise the importance of such systems, we also accept that any moderation of online content is a threat to free speech. Offensive language datasets or automatic systems can be misused to stifle disagreeing voices. Our intent is solely to learn more about the use of offensive language, learn about the various degrees of offensive language, explore how computational models can be enabled to watch and contain offensive language, and encourage others to do so. We follow the format provided by Bender and Friedman (2018) to discuss the ethical considerations for our dataset.

Institutional Review: This research was funded by the Facebook Online Safety Benchmark Research award. The primary objective of this research award is the creation of publicly available benchmarks to improve online safety. This award does not directly benefit Facebook in any way. This research was reviewed by Facebook for various aspects, in particular:

- Legal Review: Evaluates whether the research to be undertaken or the research performed can violate intellectual property rights.
- Policy and Ethics Review: Evaluates whether the research to be undertaken aligns with the best ethics practices. This includes several aspects such as mitigating harm to people involved, improving data privacy, and informed consent.

Data Redistribution / User Privacy: We extracted our data from the Pushshift Reddit dataset made publicly available by Baumgartner et al. (2020) for research purposes. The creators of the Pushshift Reddit dataset have provisions to delete comments from their dataset upon user's request. We release data in a manner that is GDPR compliant. We do not provide any user-specific

information. We release only the comment IDs and post IDs. Reddit's Terms of Service do not prohibit the distribution of ids.[8] The researchers using the dataset need to retrieve the data using the Reddit API.

Speaker and Annotator Demographic: No specific speaker demographic information is available for the comments included in Ruddit. According to the October 2020 survey published by Statista (Statista, 2020a), 50% of the Reddit's desktop traffic is from the United States. They also state that from the internet users in the US, 21% from ages 18-24, 23% from ages 25-29 and 14% from ages 30-49 use Reddit.

We restricted annotators to those residing in the US. A total of 725 crowd-workers participated in the task. Apart from the country of residence, no other information is known about the annotators. The annotators are governed by AMT's privacy policy.[9] Pew Research Center conducted a demographic survey of AMT workers in 2016. In this survey, 3370 workers participated. They found out that 80% of the crowd-workers on AMT are from the US (PRC, 2020). More information about the workers who participated in their survey can be found in their article.

It is important to include the opinions of targeted minorities and marginalized groups when dealing with the annotation of offensive language (Kiritchenko and Nejadgholi, 2020; Blackwell et al., 2017). However, we did not have our data annotated by the specific target demographic because it poses certain challenges. For example: identification of the target of offensive language; finding people of the target demographic group who are willing to annotate offensive language; and others. Annotating such offensive data can be even more traumatizing for the members of the targeted minorities. Finally, Ruddit was created with the intention to look at wide ranging offensive language of various degrees as opposed to detecting offensive language towards specific target groups.

Annotation Guidelines: We created our annotation guidelines drawing inspiration from the community standards set for offensive language on several social media platforms. These standards are made after thorough research and feedback from the community. However, we are aware that the definitions in our guidelines are not representative of all possible perspectives. The degree of offensiveness scores that we provide in Ruddit are a representation of what the majority of our annotators think. We would like to emphasize that the scores provided are not the "correct" or the only appropriate value of offensiveness. Different individuals and demographic groups may find the same comment to be more or less offensive than the scores provided.

Impact on Annotators: Annotation of harsh and offensive language might impact the mental health of the annotators negatively (Vidgen et al., 2019; Roberts, 2016, 2019; Kiritchenko and Nejadgholi, 2020). The following minimized negative mental impact on the annotators participating in our task:

- The comments that we included in our dataset are pre-moderated by Reddit's admins and subreddit specific moderators. Any comments that do not comply with Reddit's content policy are not included.[10]
- Our goal was to annotate posts one sees on social media (after content moderation). Unlike some past work, we do not limit the data to include only negative comments. We included a large sample of posts that one normally sees on social media, and annotated it for degree of supportiveness or degree of offensiveness.
- AMT provides a checkbox where requesters can indicate that some content in the task may be offensive. These tasks are not shown to annotators who have specified so in their profile. We used the checkbox to indicate that this task has offensive content.
- We explicitly warned the annotators about the content of annotation, and advised worker discretion.
- We provided detailed annotation instructions and informed the annotators about how the annotations for offensive language will be used for studying and understanding offensive language.
- The annotation of our data was crowdsourced, allowing for a large number of raters (725). This reduces the number of comments seen per rater. We also placed a limit on how many posts one may annotate. Annotators were not allowed to submit more than $\sim 5\%$ of the total assignments.
- There are just 25 comments in the top 10% of the offensiveness score range. Thus, most annotators ($> 99.95\%$) do not see even one such comment.

[8] https://www.reddit.com/wiki/api-terms
[9] https://www.mturk.com/help
[10] https://www.redditinc.com/policies/content-policy

Identity Terms: As discussed in section 5, in Ruddit, certain identity terms show a higher association with offensive comments than with the supportive comments. In order to address this, we created a variant of Ruddit, in which we replaced all the identity terms (from the list given in Appendix A.4) with the [*group*] token. We call this variant the *identity-agnostic* dataset. We release the code for creating this variant from the original dataset. We evaluated our computational models on this variant and observed that the models did not learn to benefit from the association of the identity terms with the offensive comments.

Computational Models: The models reported in this paper are not intended to fully automate offensive content moderation or to make judgements about specific individuals. Owing to privacy concerns, we do not model user history to predict offensiveness scores (Mitchell et al., 2018).

Feedback: We are aware that our dataset is subject to the inherent bias of the data, the sampling procedure and the opinion of the annotators who annotated it. Finally, we acknowledge that this is not a comprehensive listing of all the ethical considerations and limitations. We welcome feedback from the research community and anyone using our dataset.

References

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020a. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020b. Hatebert: Retraining bert for abusive language detection in english.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*,

Chapters, chapter 8, pages 178–201. Edward Elgar Publishing.

P. Fortuna and S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *CoRR*, abs/1802.00393.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266.

Timothy Jay and Kristin Janschewitz. 2008. *The pragmatics of swearing*. Walter de Gruyter GmbH & Co. KG.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia. Association for Computational Linguistics.

David A. Jurgens, Peter D. Turney, Saif M. Mohammad, and Keith J. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, page 356–364, USA. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko and Isar Nejadgholi. 2020. Towards ethics by design in online abusive content detection.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. 2014. Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

J. J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. working paper.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Scott C. Lowe. 2016. Stratified validation splits for regression problems.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model cards for model reporting. *CoRR*, abs/1810.03993.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Emily Munro. 2011. The protection of children online: A brief scoping review to identify vulnerable groups.

B. Orme. 2009. Maxdiff analysis: Simple counting,individual-level logit, and hb. sawtooth software, inc.

C.E. Osgood, G.J. Suci, and P.H. Tenenbaum. 1957. *The Measurement of meaning*. University of Illinois Press, Urbana:.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances.

PRC. Research in the crowdsourcing age, a case study [online]. 2020. Accessed: 2020-10-10.

Stanley Presser and Howard Schuman. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc.

Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online. Association for Computational Linguistics.

Sarah T. Roberts. 2016. *Chapter Eight: Commercial Content Moderation: Digital Laborers' Dirty Work*. Peter Lang.

Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

James Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110:145–72.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

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

Wiktor Soral, M. Bilewicz, and M. Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44:13V–146.

Statista. Regional distribution of desktop traffic to reddit.com as of september 2020, by country [online]. 2020a. Accessed: 2021-01-04.

Statista. Share of adult internet users in the united states who have personally experienced online harassment as of january 2020 [online]. 2020b.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

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 & Social Computing*, CSCW '14, page 415–425, New York, NY, USA. Association for Computing Machinery.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Gregor Wiedemann, Eugen Ruppert, Raghav Jindal, and Chris Biemann. 2018. Transfer learning from LDA to bilstm-cnn for offensive language detection in twitter. *CoRR*, abs/1811.02906.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Wikipedia Talk Corpus.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A Supplemental Material

A.1 Post and Comment Criteria

We selected the posts from the subreddits based on the following criteria:

1. **Date**: To extract comments from posts that discuss current matters, we took comments from the time period of January, 2015 to September, 2019 (last available month at the time of extraction).

2. **Thread length**: We chose posts with more than 150 comments and less than 5000 comments. This criteria ensured that the posts contained enough comments to capture meaningful discussion.

3. **Post length**: We chose posts containing more than 5 words and less than 60 words in the post body. This was done to avoid posts that are too short to provide enough information or are too long and have a possibility of being spam.

4. **URL**: Often, posts on Reddit contain URLs redirecting to images, videos, news articles and others. We limited our posts to those containing at most one URL to avoid issues arising due to missing context.

For each post, the hierarchical threads were reconstructed using the *Anytree* python library. We filtered comments from these posts based on the following criteria:

1. **Comment length**: We chose comments containing more than 5 words and less than 150 words in the comment body. We did this to include comments that are neither too long (can be difficult to annotate) nor too short (not very valuable).

2. **No. of users**: In the first and last 25 comments of the thread, we ensured participation of at least 4 users. This was done to ensure that the comments in our dataset are from a diverse set of users.

3. **URL**: We chose comments with no URL in them. Comments with URL can be difficult to annotate as the URLs provide extra context for the comment.

A.2 Annotation

Figure 5 shows the detailed annotation instructions given to the crowd-workers for the task.

A sample questionnaire for the final annotation task is shown in Figure 6.

The hourly compensation rate for annotators on Amazon Mechanical Turk was US$7.50/hr. The task received considerable attention with 725 participants in total.

A.3 Sample data

Table 6 contains comments from Ruddit grouped according to the 5 score bins.

A.4 Data Analysis

We used the list of identity terms used by Dixon et al. (2018) with a few of our own additions. The terms used are *lesbian, gay, bisexual, transgender, trans, queer, lgbt, lgbtq, homosexual, straight, heterosexual, male, female, nonbinary, african, africanamerican, black, white, european, hispanic, latino, latina, latinx, mexican, canadian, american, asian, indian, middle eastern, chinese, japanese, christian, muslim, jewish, buddhist, catholic, protestant, sikh, taoist, old, older, young, younger, teenage, millenial, middle aged, elderly, blind, deaf, paralyzed, atheist, feminist, islam, muslim, man, woman, boy, girl.*

Figure 7 shows a histogram of the comments containing swear words–degree of offensiveness, over 40 equi-spaced score bins of size 0.05.

Figure 8 shows a distribution of comments within each of the 5 score bins over the subreddits that were included in the Topics category.

A.5 Computational Modeling

Hyperparameter Tuning We tuned hyperparameters for the BERT and the BiLSTM models. We performed grid search cross-validation on Ruddit and used Pearson's r to select the best hyperparameter setting. All experiments were performed on a fixed seed value of 12.

For the BiLSTM model, the batch size was fixed at 32 and the number of epochs was set to 7. The hyperparameter search space is as follows:

- Number of Layers (N): $1, 2$
- Hidden size (H): $64, 128, 256$

For the BERT model, the batch size was fixed at 16 and BERT tokenizer's maximum length was set

Social media conversations often include an original post followed by other comments in response to the original post. These response comments can vary from being highly supportive and pleasant to highly offensive and abusive. We are interested in determining the degree of offensiveness/supportiveness of response comments. Specifically, in each HIT, you will be given four comments. Your task is to mark:

- the comment that is the MOST offensive (LEAST supportive).
- the comment that is the LEAST offensive (MOST supportive).

Purpose of the task:

Your responses will be used to better understand the type, nature and degree of offensive and supportive language.

Notes:

- The comments included in this task may contain foul, abusive language (including but not limited to sexist, violent, vulgar or racist terms).
- Try not to overthink the answer. Let your instinct guide you.
- There are a certain number of gold annotations interspersed throughout the data. If you mark these gold questions incorrectly, we will give you immediate feedback in a pop-up box. We will keep track of your answers for these gold questions. If you mark too many of these incorrectly, it might lead to rejection of your HITs, upon further scrutiny.

Instructions:

- For the MOST offensive (LEAST supportive) question: If all four comments are not offensive, then select the comment that is most negative or least supportive. If two (or more) comments are equally offensive, and both are MOST offensive (among the four options), then choose any one of these equally offensive comments as the MOST offensive comment.
- For the LEAST offensive (MOST supportive) question: If all four comments are not supportive, then select the comment that is most positive or least offensive. If two (or more) comments are equally supportive, and both are MOST supportive (among the four options), then choose any one of these equally supportive comments as the MOST supportive comment.
- Consider 'offensive comments' to be a broad category that includes abhorrent, abusive, rude, objectionable, disagreeable, and discourteous comments. Offensive comments often denigrate, criticize, hurt, or harm the perception of self of another person or group (with or without using offensive words). Offensive comments tend to include aggressive outbursts, name calling, anger, hostility, bullying, or hurtful sarcasm, but they can be more subtle as well.
- Consider 'supportive comments' to be a broad category that includes agreeable, delightful, kind, nice, pleasant, pleasing, respectful, and soothing comments.
- You cannot pick the same comment for both the categories.
- Responses that are not in accordance with the instructions will not be paid for.

Figure 5: Detailed instructions for the final annotation task.

Q. From the four comments below, choose the comment which is **LEAST** OFFENSIVE (most supportive) and the comment which is **MOST** OFFENSIVE (least supportive).

Least Offensive	Comment	Most Offensive
○	The first time I had sex the guy took off the condom and didn't tell me. Men do this shit too. I don't care about how you feel about it, shitty people exist in both genders.	○
○	Promo girls is just a euphemism for escort right Cos I m seeing a theme tonight	○
○	I've heard of quite a few guys having sex with girls who were lying about being on birth control, and had to live with it for the rest of their lives. This does happen quite often	○
○	I'm clearly not as good at this as I think I am. That being said it seems like there's a very intentional effort to conflate (a) "I was just sitting in a known gang area and a fight broke out and I got hit" with (b) a woman was dragged behind a dumpster and raped because (a) is more common than (b) which obscures the fact that (b) happens almost exclusively to women.	○

Figure 6: Sample questionnaire for the final annotation task.

to 200. We tune hyperparameters on the settings that Devlin et al. (2019) found to work best on all tasks. The search space is as follows:

- Learning rate: $2e-5$, $3e-5$, $5e-5$
- Number of epochs: $3, 4$

We reported the best setting for the models in section 6.1. The average r of the BERT and the BiLSTM models across all hyperparameter search

Bin	Comment	Score
1	Truly gave me a smile	−0.812
	Awwwwwe That's was the best reveal I've seen.	−0.688
	I'm literally doing the same tonight!	−0.625
	Have an awesome time :)	
2	My apologies for missing that in your post. Carry on.	−0.583
	I'm assuming this is a great response for i haven't seen the video in question, but i do believe that the woman in the video carried herself very elegantly!	−0.375
	Could you not love a "purrfectly" healthy moggy? One whose life you could save from a shelter and won't break the bank.	−0.25
3	Handwritten letters are pure gold. It is a simple but immensely pleasant gift to receive. Still have the ones my ex gf would sent me when we were apart.	−0.196
	It would take way more time to get a screaming baby to stay still long enough to inject them. I remember my little sister throwing off doctors and nurses like a tiny she-hulk when she was a toddler. She also punched my dad in the face.	−0.125
	Oh man, that's a great idea! Nuke a few large cities in the US, this will send the economy to the sky!	0.104
4	No it's usually their kids they are killing.	0.208
	They were compassionately raped too. That's how you feel the love.	0.312
	You look like a lesbian mechanic who has a shell collection	0.438
	You know it sucks when you cough, and the all of a sudden you got whooping gay	0.457
	The best thing about anti-vax is soon we wont have to do anything about them, the diseases will do it for us	0.458
5	She's a hoe, plain and simple.	0.625
	This guy is giving me fucking aids	0.792
	fuck you, you're just pretending to be blocked. get back to work RIGHT NOW you piece of shit piping	0.938

Table 6: More sample comments from Ruddit for each of the 5 score bins. Comments in bold are implicitly offensive.

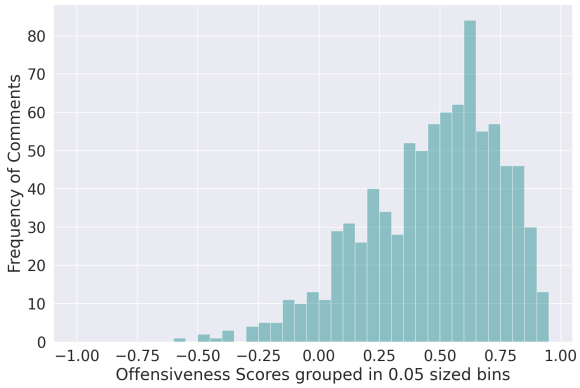

Figure 7: A histogram of frequency of comments containing swear words–degree of offensiveness. Degree of offensiveness scores are grouped in bins of size 0.05.

trials was 0.868 ± 0.005 and 0.827 ± 0.002 respectively.

Training Times We trained all our models on the Tesla T4 GPU. The number of GPU(s) used is 1. The number of trainable parameters and thus, the training time varied for each model. The approximate number of trainable parameters for each model is as follows:

- BiLSTM ($N = 2$, $H = 256$): 7 million
- BERT: 108 million
- HateBERT: 109 million

The approximate average runtime for each model on the Ruddit dataset is as follows:

- BiLSTM ($N = 2$, $H = 256$): 2 seconds per epoch
- BERT: 3 minutes per epoch
- HateBERT: 3.6 minutes per epoch

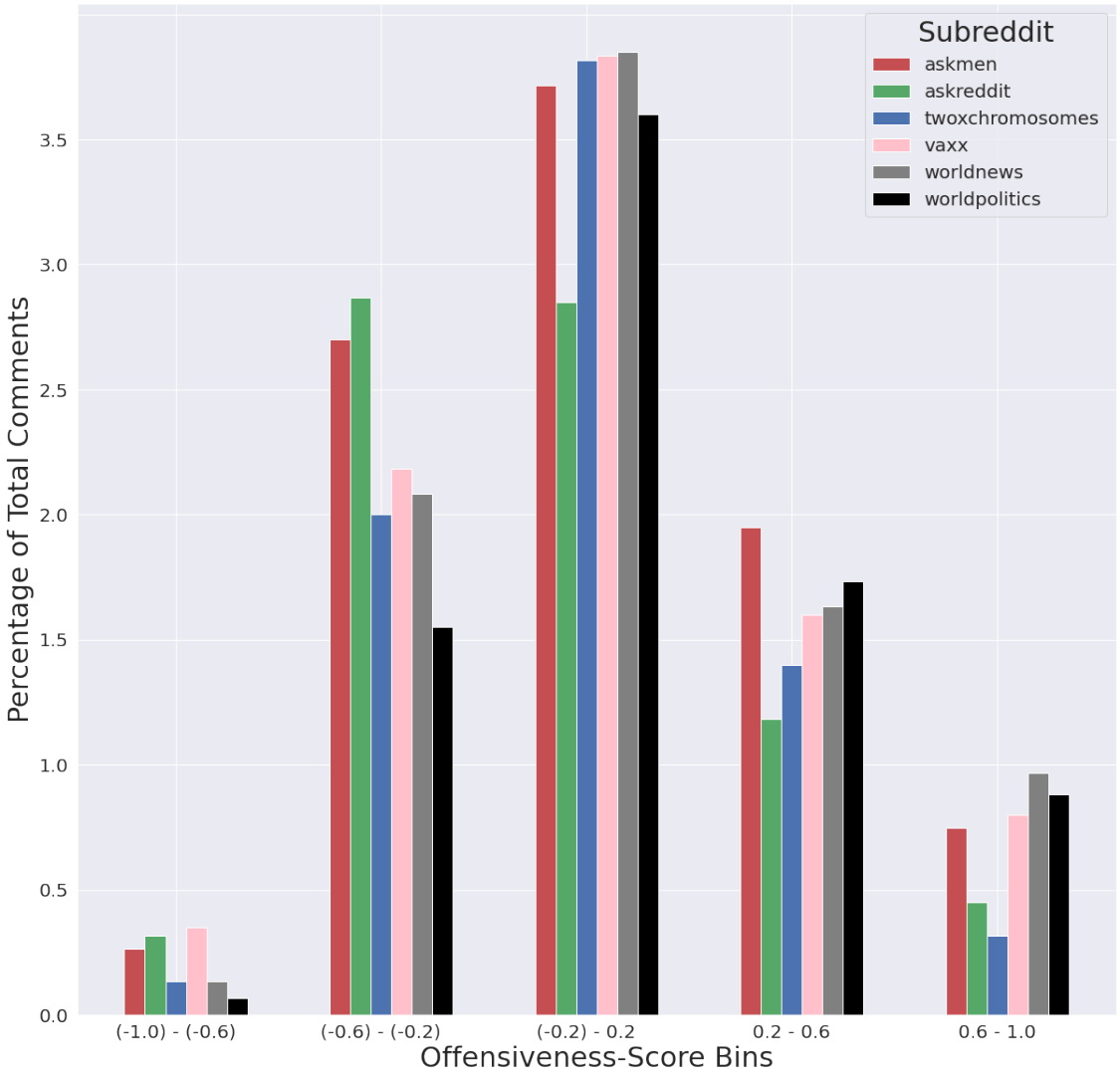

Figure 8: Distribution of comments from each subreddit in each offensiveness-score bin.