Abusive Language Detection in Heterogeneous Contexts: Dataset Collection and the Role of Supervised Attention

Hongyu Gong,[1] **Alberto Valido,**[2] **Katherine M. Ingram,**[2]
Giulia Fanti,[3] **Suma Bhat,**[1] **Dorothy L. Espelage**[2]

[1] University of Illinois at Urbana-Champaign, [2] University of North Carolina at Chapel Hill, [3] Carnegie Mellon University
hgong6@illinois.edu, avalia@unc.edu, ingramkm@live.unc.edu,
gfanti@andrew.cmu.edu, spbhat2@illinois.edu, espelage@unc.edu

Abstract

Abusive language is a massive problem in online social platforms. Existing abusive language detection techniques are particularly ill-suited to comments containing *heterogeneous* abusive language patterns, i.e., both abusive and non-abusive parts. This is due in part to the lack of datasets that explicitly annotate heterogeneity in abusive language. We tackle this challenge by providing an annotated dataset of abusive language in over 11,000 comments from YouTube. We account for heterogeneity in this dataset by separately annotating both the comment as a whole and the individual sentences that comprise each comment. We then propose an algorithm that uses a supervised attention mechanism to detect and categorize abusive content using multi-task learning. We empirically demonstrate the challenges of using traditional techniques on heterogeneous content and the comparative gains in performance of the proposed approach over state-of-the-art methods.

Introduction

Abusive language refers to strongly impolite and harmful language used to hurt and control a person or a group by way of harassment, insults, threats, bullying and/or trolling (Waseem et al. 2017). Because of the nature of digital technology, abusive language can be generated anonymously and can spread to many victims in a short time, making it a serious societal concern (Price, Dalgleish et al. 2010). Due to the profound negative impact of abusive language, many online platforms today dedicate significant resources to its detection and categorization (Nobata et al. 2016). As the problem grows in scope and scale, so does the need for automated detection and categorization tools.

Despite significant prior work on the automated detection of abusive language, it remains a difficult task (Vidgen et al. 2019). One important reason for this difficulty is *heterogeneity* of abuse: abusive comments often contain a combination of abusive and non-abusive language, and it can be difficult for algorithmic approaches to understand this distinction. Table 1 illustrates examples of heterogeneous abusive comments from YouTube. By *sentence-level heterogeneity*, we mean multi-sentence comments where some sentences are abusive, and others are not. Comment 1 shows an instance where removing the abusive sentence does not affect the meaning of the comment. Comment 2 is more subtle; removing the underlined sentence completely changes the comment's meaning. By *phrase-level heterogeneity*, we mean comments where only a few words are abusive, but there are no abusive full sentences.[1] Comment 4 illustrates a combination of both types of heterogeneity. Notably, even though the comment is abusive, it is also pro-social, in that it is defending the victim of another abusive comment. Hence, there are substantial subtleties associated with detecting and understanding heterogeneous abusive language—both for humans and for automated detectors.

Sentence-level heterogeneity	1	This case really exposes Sheehy as a total fraud. **Not to mention a total man hating cunt.** You have done a masterful (mistressful?) job of exposing this lying fraudster. Love the ending when George is cited as a MGTOW Hero! He is indeed.
	2	+Username I understand. Personally, I rarely do jokes involving violence of any sort. **(Unless it involves Muslims!)** :)
Phrase-level heterogeneity	3	+Username **I always hated him**, cause he makes his opinion sound like its all facts plus that and **he sounds like a fuckin' faggot**
Both	4	+Username **Shut the fuck up, cunt.** I'm a man, and I respect her a lot, because she speaks truth. She is very eloquent and a pleasure to look at. **You fucking idiot,** all you do is make fun of peoples looks, did you miss the points she was making?

Table 1: Examples of YouTube comments with heterogeneous abusive language. Abusive parts are underlined in bold text.

Heterogeneity does not necessarily lessen the effects of abusive language on victims. Both cyber- and traditional bullies are known to sometimes engage in a combination of friendly and bullying behaviors, while still negatively affecting victims (Kowalski and Limber 2007; James et al. 2011). Similarly, workers who receive a combination of positive and negative feedback tend to experience stronger overall negative emotional reactions (Choi et al. 2018). Hence, it is

[1] Considering a comment with more than one independent clause, we define it as being "phrase-level heterogeneous" if it contains at least one independent clause that is non-abusive even though the comment as a whole is labelled as abusive, or vice versa.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

important for automated abusive language detectors to identify abusive language couched in non-abusive language (and vice versa).

Unfortunately, detection algorithms today struggle with heterogeneity. The state-of-the-art approach for automated abusive language detection relies on supervised methods that predict abusive language by learning a stacked bidirectional LSTM with attention (Chakrabarty, Gupta, and Muresan 2019). However, RNNs are known to perform poorly on lengthy and/or heterogeneous data in other domains (Wang 2018). Indeed, we also find that such techniques have poor detection performance on heterogeneous comments because they struggle to identify which *part* of an abusive comment makes it abusive (e.g., Fig. 2).

The challenges are caused in part by a dearth of datasets explicitly annotating heterogeneity of abuse. Recently Vidgen and Derczynski (Vidgen et al. 2019) categorize existing abusive language datasets from the natural language processing (NLP) literature, and show that the majority of existing datasets (27/50) are collected from Twitter (e.g., (Waseem and Hovy 2016; Davidson et al. 2017; Golbeck et al. 2017)), which imposes a hard limit of 280 characters per tweet. We hypothesize that heterogeneity may be less common in such short comments. Regardless, these datasets are annotated per tweet, and therefore cannot capture heterogeneity of abuse within tweets. Although there are datasets with longer abusive comments (Qian et al. 2019; Ribeiro et al. 2021), they are also typically annotated at the comment level [2].

Because of this data availability problem, abusive language detection algorithms have been tuned (and possibly overfitted) to microblogging platforms with short comments. At the same time, many platforms, such as Facebook and YouTube, attract longer comments that naturally include a broader range of abusive language patterns of interest, including heterogeneity (Schmidt and Wiegand 2017).[3]

To summarize, heterogeneous abusive language detection is difficult today because: (a) there is little labelled data in this category, and (b) even with labelled data, existing techniques are not well-suited to heterogeneous comments. In this work, we make three contributions:

(1) We provide the first annotated dataset of over 11,000 comments in English collected from over 250 YouTube channels related to feminism. The annotation was performed using a theoretically-grounded abusive language taxonomy. In addition to annotating each comment as abusive language or not, we also provide sentence-level annotations to study the role of heterogeneity. Abusive comments are categorized into types of group-directed abuse; this is a currently under-explored area owing to lack of datasets (Fortuna and Nunes 2018).

(2) Using our YouTube dataset, we demonstrate the challenges associated with using traditional abusive language techniques on heterogeneous content.

(3) We propose a model with supervised attention mechanism for detecting abusive language and evaluate it on our dataset. This novel attention mechanism emulates the human decision-making process in the presence of heterogeneity. Toward this, our novel attention encoder[4] maps the real-valued model attention and binary human attention to the same space.

(4) We show that such an explicit supervision of attention results in gains of over 2% in abusive language detection ROC AUC over the best competing baseline, and similar gains on an abusive language categorization task, which aims to classify the *nature* of abusive language. Additionally, although our YouTube dataset only contains annotations of sentence-level heterogeneity, we find that our approach also improves detection in instances with phrase-level heterogeneity.

Related Work

Abusive language detection. Automated abusive language detectors range from supervised machine learning models built using a combination of manually crafted features such as n-grams (Wulczyn, Thain, and Dixon 2017), syntactic features (Nobata et al. 2016), and linguistic features (Yin et al. 2009; Joksimovic et al. 2019), to more recent neural networks (Park and Fung 2017; Maity et al. 2018). The most recent studies on abuse detection have reported state-of-the-art performance using RNNs with the attention mechanism (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017; Chakrabarty, Gupta, and Muresan 2019). Challenges to abusive language detection include the linguistic variety and nuances (Nobata et al. 2016; Schmidt and Wiegand 2017), and the inherent biases in dataset creation (Vidgen et al. 2019).

Abusive language categorization. There has been a lack of abusive language datasets with finer-grained taxonomies (Park and Fung 2017; Badjatiya et al. 2017; Fortuna and Nunes 2018). A fine-grained abusive language classification provides insights into the nature of abusive language, permitting a more targeted mechanism for detection and intervention (Hoff and Mitchell 2009). Common approaches to categorization have been learning a separate classifier for each category and relying on feature engineering (Van Hee et al. 2015; Dinakar, Reichart, and Lieberman 2011). In this study, we extend prior work by creating a dataset using a taxonomy of 4 abusive categories and then using it to train a neural categorization model via multi-task learning, an approach not explored in prior work.

Attention-based models. The attention mechanism is widely incorporated into neural networks to identify focus regions in inputs (e.g., when the decision hinges on the presence of key phrases). Combined with LSTMs and learned in an unsupervised manner, attention was found to help models achieve good performance in certain NLP applications, e.g., (Luong, Pham, and Manning 2015). Recent works have also explored the use of added supervision on the attention mechanism and found it to help machine translation with annotated alignment information (Liu et al. 2016), event detection with annotated arguments (Liu et al. 2017) and domain transfer with human rationale (Bao et al. 2018).

[2]To the best of our knowledge, the only dataset with sentence-level annotation is (de Gibert et al. 2018) with the group-directed abusive content limited to hate speech. This dataset also does not include a separate annotation for the comment as a whole.

[3]Although there exist several datasets for Facebook, they are either not in English and/or synthetic data (Chung et al. 2019).

[4]We release the data and the code at https://github.com/HongyuGong/Abusive-Language-Detection-Categorization.

Our approach is to use human rationale of abuse detection towards training robust and interpretable models with supervised attention.

Dataset and Annotation

Our objective is to study abusive language detection in heterogeneous settings, where individual comments may occur as a combination of sentences with abusive and non-abusive language, illustrated via representative examples in 1. Existing datasets are ill-suited to this task for two reasons: (1) They generally consist of comments that are too short to observe heterogeneity. (2) Even in longer comments that exhibit a mixture of abusive and non-abusive language, existing datasets do not include annotations that highlight the specific *portion* of each comment that is abusive. Our goal was to build a dataset that addresses both of these problems, through a process that was largely consistent with the recommendations of (Vidgen et al. 2019).[5]

Data Collection

We chose to study the YouTube platform, where comments tend to be longer. Despite being an open platform, the only public dataset of abusive language from YouTube is in Arabic (Alakrot, Murray, and Nikolov 2018). We collected a total of $11,540$ public comments posted on 253 YouTube channels as of May 2017, with an average comment length of 32 words. These channels were selected as the top results for the keyword *feminism*. We used videos related to feminism for two reasons: (1) we observed a high occurrence of abusive language in the results, which helped us isolate the effects of class imbalance. (2) Due to IRB restrictions, we needed a channel with limited presence of children. We used an independent service (socialbook.io), which estimated 89% of commenters on these channels were above 17 years of age. Our choice to focus on feminism-related videos was made purely for pragmatic reasons, and does introduce bias to the dataset. However, one secondary benefit may be that it can aid concurrent efforts to understand the "manosphere" (Ribeiro et al. 2021).

Annotation Process

Abusive language detection models trained on data annotated by experts have better performance and generalization (Waseem 2016); hence, our dataset was annotated by a diverse team of 17 psychology students, of whom 3 research coordinators were graduate students studying bullying and related phenomena. Per the recommendations of (Vidgen et al. 2019), the annotators and coordinators represented a range of ethnicities, genders, and mother tongues.

We began the process with a training session, in which annotators were given our definition of abusive language, as well as examples of (non-)abusive language, including borderline cases (Vidgen et al. 2019). Because multiple definitions of abusive language are available (Fortuna and Nunes 2018; Peter and Petermann 2018), we chose the definition of abusive language to be "an expression that is intended to hurt or attack an individual or a group of people on the basis of race, appearance, gender identity, religion, or ethnicity/nationality". This definition was based on related literature (e.g., (Nobata et al. 2016)) and naturally overlaps with the notions of profanity and hate speech (Nockleby 2000). Apart from this high-level definition of abusive language, we developed a codebook for a taxonomy of group-directed abusive language that was theoretically informed by a synthesis paper (Patchin and Hinduja 2015).

Once the initial training was complete, we began a three-phase annotation process. *(1) Comment-level labeling*: Each comment was first classified as either abusive or non-abusive by assigning it to an annotator pair, each of whom would annotate the comment independently. Then, the annotators discussed any disagreements in the annotation and came to consensus. Noting the importance of context in annotation, annotators had access to the comments that preceded any given comment; this was a significant departure from previously-collected datasets (Vidgen et al. 2019). In all, 27.5% of comments were classified as abusive.

(2) Sentence-level labeling: Next, the comments were split into a total of $26,373$ sentences. Each sentence was individually labeled as either abusive or non-abusive (even if the comment had previously been labeled as non-abusive) in the context of the comment. This nested annotation provided the valuable localized human rationale for supervised attention training in our model. It also provided crucial insights about the nature of heterogeneity in the dataset; we found that in a multi-sentence comment, one sentence being labelled as abusive generally caused the comment to be classified as abusive. In fact, there were examples of comments with abusive language being used to defend victims of other abusive language (see Table 1 for examples). Among abusive comments, 43.4% are a mixture of abusive and non-abusive sentences. Including phrase-level annotations would have increased the granularity of heterogeneity, but we focused on identifying only sentence-level heterogeneity to reduce the annotation effort of an already difficult and time-consuming process.

(3) Comment categorization: Abusive comments were further classified into four predefined content categories, an aspect that is not available in a vast majority of datasets. Similar to (Founta et al. 2018), our annotation scheme considered four categories of group-directed abusive language: (a) gender and sexuality, (b) race, nationality and ethnicity, (c) appearance and individual characteristics, and (d) ideology, religion or political affiliation. Each sentence and comment was annotated by two annotators at each stage, and the inter-annotator agreement (Cohen's κ) was 88% for the abusive sentence labeling task, 90% for the abusive comment labelling, and 93% for the categorization task. Annotators later met to resolve their differences. A total of 1979 abusive comments were labelled into these categories.

Group meetings We led weekly annotation team meetings to discuss disagreements in annotations between each pair of annotators. The team would come to consensus as a group by consulting third-party resources (e.g., Urban Dictionary) and by relying on the diverse cultural contexts of the group. These meetings had several benefits:

(1) Annotator subjectivity is known to be a major factor af-

[5]Usernames were anonymized by a generic token.

fecting the quality of labels for machine learning pipelines (Hube, Fetahu, and Gadiraju 2019). Discussing difficult-to-classify comments, particularly among a culturally-diverse group of annotators, was therefore important for ensuring the quality of our annotations (Vidgen et al. 2019).

(2) Miceli *et al.* recently demonstrated that annotators tend to view the opinions of their supervisors as authoritative, and defer to their judgment (Miceli, Schuessler, and Yang 2020). We explicitly aimed to avoid such an effect by having annotators guide *each other* to consensus.

(3) Our weekly meetings were also designed to monitor the emotional health of our annotators. Abusive language is known to have negative psychological effects on bystanders as well as victims (Low et al. 2007; Ferguson and Barry 2011). We observed such effects among our research assistants, who described the experience as "a taxing process, psychologically" and the content as "appalling." We therefore discussed the annotators' emotional state at each meeting.

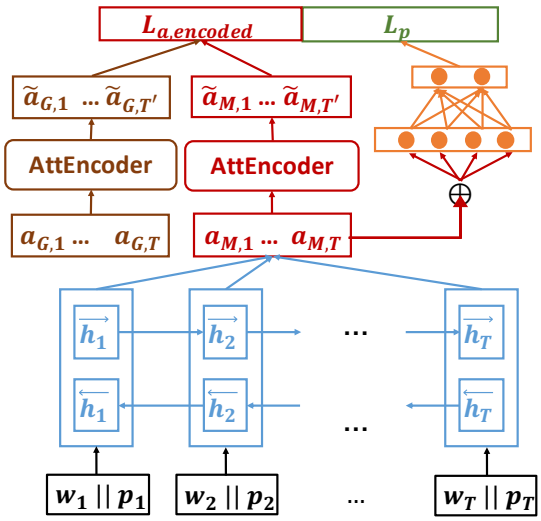


Figure 1: RNN with supervised attention. Each input word is a concatenation of its word- and POS embeddings. The input is a sequence of embeddings, and the recurrent layer generates hidden state vectors. The attention module determines the attention distribution over these hidden vectors, which are linearly weighted with the attention weights to input to the feedforward network (FFN). The FFN predicts the scores of each class, which are transformed to a probability distribution over the classes in the output. The AttEncoder maps the ground truth (G) and model (M) attention to the same vector space in order to measure the encoded attention loss.

Abusive Language Detection

About 43.4% of the abusive comments in our YouTube data are a mixture of both abusive and non-abusive content. We hypothesize that manually demarcating the abusive portion in a comment from other non-abusive content (as a way of showing the human rationale) can provide better supervision while training abusive language detectors.

Attention-based models available in existing literature train the attention component implicitly since only the predictive function (and not the attention mechanism) is supervised. Instead, we conjecture that by explicitly supervising the attention mechanism we might be able to steer the model towards learning the relevant patterns based on human rationale. For example, given a comment "+Username don't fret, bearing, all know you're a cunt and a right excellent one at that", we visualize the (implicitly trained) attention weights of a Recurrent Neural Network (RNN) over all the words as shown in Fig. 2(a). We see that the model classified the comment as abusive because it wrongly considered *fret* as a signal of abuse instead of *cunt*.

Our goal in this study is to introduce the idea of supervised attention and train the neural network not only to give correct predictions but also to accurately identify abusive patterns from the input. We will show next that supervised attention is beneficial to both abusive language detection and categorization.

RNN with Supervised Attention

In this study, we use a Recurrent Neural Network with bidirectional LSTM units (a BiLSTM network) owing to its state-of-the-art performance in abusive language classification (Chakrabarty, Gupta, and Muresan 2019). Let $\{w_1, w_2, \ldots, w_T\}$ be the T words of a comment. Inspired by factored neural network models which incorporate extra-linguistic information, we use the part-of-speech (POS) tags of the words $\{p_1, p_2, \ldots, p_T\}$ in addition to the words as the input sentence (Sennrich and Haddow 2016). For every input word w_t, we concatenate its word embedding and POS embedding as its vector representation $\mathbf{x}_t$. The word and POS embeddings are pretrained with the word2vec CBOW model on the training data (Mikolov et al. 2013). The structure of our RNN model with attention supervision is shown in Fig. 1.

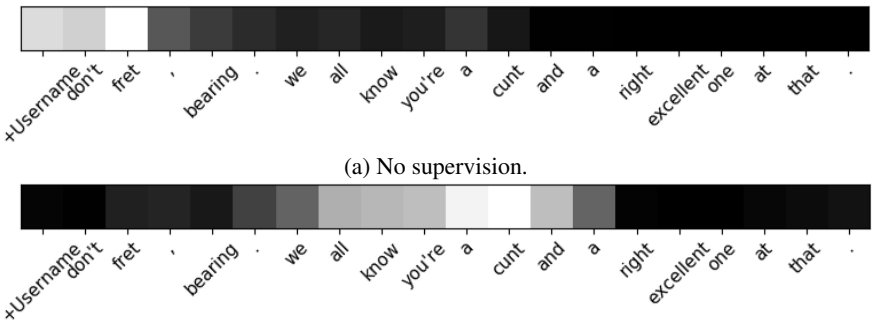
The BiLSTM recurrent layer incorporates contextual information from both sides of a given word; it concatenates the two state vectors $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ to yield the bidirectional vector $\mathbf{h}_t^b$, i.e., $\mathbf{h}_t^b = [\overrightarrow{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t]$. We apply the attention mechanism proposed in (Yang et al. 2016) and also used in (Chakrabarty, Gupta, and Muresan 2019). The model's attention value $a_{M,t}$ on the t-th token is calculated from hidden state vector $\mathbf{h}_t^b$:

$$\mathbf{v}_t = \sigma(\mathbf{W}_u \mathbf{h}_t^b + \mathbf{b}_u), \tag{1}$$

$$a_{M,t} = \frac{\exp(\mathbf{u}^T \mathbf{v}_t)}{\sum_{t'} \exp(\mathbf{u}^T \mathbf{v}_{t'})}, \tag{2}$$

where $\mathbf{W}_u$ is a trainable matrix, $\mathbf{b}_u$ and $\mathbf{u}$ are vectors in the recurrent layer, and $\sigma(\cdot)$ is the sigmoid function. The hidden state vectors at different time steps are linearly weighted to yield the vector $\mathbf{z} = \sum_t a_{M,t} \mathbf{h}_t^b$, a compressed representation of the input sentence. We have two FFN layers with the sigmoid activation and an output layer above the recurrent layer, which outputs $\mathbf{y}'$ as the model's predicted probability distribution over the classes.

The cross-entropy prediction loss L_p is used to measure the difference between the predicted vector $\mathbf{y}'$ and the ground truth vector $\mathbf{y}$, where $\mathbf{y}'$ and $\mathbf{y}$ are two-dimensional vectors for the binary abusive classification task.



(a) No supervision.

(b) Encoded supervision.

Figure 2: Attention visualization of RNN trained with and without attention supervision for an abusive sentence in grey-scale image. The lighter the word, the higher its attention weight.

To supervise the attention over the input sentences and evaluate how well the neural network can target the truly abusive segments, we define the attention loss L_a as part of our training objective. To indicate the abusive part in a comment, we assign a (ground-truth) binary attention vector $\mathbf{a_G}$ to the input sequence, where the words in the abusive segment are marked as 1 and others as 0. This attention vector is derived from the sentence-level labels in our data (see the section of Dataset and Annotation) and is used to train the attention mechanism. The ground-truth attention $\mathbf{a_G}$ is scaled by $1/(\mathbf{1^T a_G})$ to normalize the weights. Without loss of generality, we continue to use $\mathbf{a_G}$ to refer to the scaled ground truth attention. We use $\mathbf{a_M}$ to refer to the attention vector output from the attention module of our system, that is based on the outputs of the RNN.

Ideally the estimated attention $\mathbf{a_M}$ should correspond to the ground truth attention $\mathbf{a_G}$, i.e., higher model attention weights should be assigned to the words that are marked with non-zero values. Toward this goal of aligning $\mathbf{a_M}$ and $\mathbf{a_G}$, we consider two simple but commonly used losses in neural network training: the L1 loss (L_{a,l_1}), and the L2 loss (L_{a,l_2}). The use of the L2 attention loss has been explored in several previous works, including (Liu et al. 2016, 2017).

Encoding attention. Because our ground truth annotation does not mark the degree to which words are abusive, all the words in an abusive segment are assigned non-zero weights. More specifically, the components of the ground truth attention vector are either zero or uniformly non-zero. In contrast, the model attention is real valued, assigning different weights to the words. As a result, we note that the attention vectors $\mathbf{a_M}$ and $\mathbf{a_G}$ are not in the same vector space, and hence computing the L1 and L2 losses over these attention vectors as is may not accurately capture their correspondence. Toward remedying this situation, we propose an attention encoder (termed as *AttEncoder*) to encode the ground truth and the model attention vectors to a common space and to then estimate how well they match. This is done via a neural module, which is shown in Fig. 1. The ground truth attention $\mathbf{a_G}$ is encoded as $\tilde{\mathbf{a_G}}$ using an AttEncoder as indicated below.

$$\tilde{\mathbf{a}}_G = \text{AttEncoder}_G(\mathbf{a}_G) = \tanh(\mathbf{W}_G\mathbf{a}_G + \mathbf{b}_G), \quad (3)$$

where $\tanh(\cdot)$ is an activation function, matrix $\mathbf{W}_G$ and bias vector $\mathbf{b}_G$ are tunable parameters.

Similarly, the model attention $\mathbf{a}_M$ is transformed by another AttEncoder:

$$\tilde{\mathbf{a}}_M = \text{AttEncoder}_M(\mathbf{a}_M) = \tanh(\mathbf{W}_M\mathbf{a}_M + \mathbf{b}_M), \quad (4)$$

where $\mathbf{W}_M$ and $\mathbf{b}_M$ are tunable parameters. The resulting attention vectors $\tilde{\mathbf{a}}_G$ and $\tilde{\mathbf{a}}_M$ are in the same hidden space and their inner product can be interpreted as their similarity. The encoded attention loss estimated by this module is $L_{a,\text{encoded}}$:

$$L_{a,\text{encoded}} = -\tilde{\mathbf{a}}_G^T\tilde{\mathbf{a}}_M. \quad (5)$$

The total loss L is defined to be a weighted sum of the prediction loss L_p and the attention loss L_a:

$$L = L_p + \beta L_a, \quad (6)$$

where L_a can be one of the three attention losses, $L_{a,1}$, $L_{a,2}$ and $L_{a,\text{encoded}}$. The hyperparameter β is tuned on the validation set, and $\beta = 0.2$ in our experiments. The neural network is trained from end to end to minimize the total loss with explicit attention supervision L_a, thereby capturing abusive patterns and making classification decisions.

Experiments on Abuse Detection

We compare our system with previous models used for abusive language detection. In our experiments, we randomly split the annotated data (at the comment-level) into training, validation and test sets in a ratio of 3:1:1 for use in the abuse detection and categorization tasks. The baselines are:

(1) Support vector machine (SVM)—included in this comparison because of its competitive performance in related prior works (Van Hee et al. 2015; Nobata et al. 2016). It takes word unigrams, bigrams and character trigrams as features. The vocabulary sizes of word- and character- ngrams are 5,000. We also use the sentiment feature from (Van Hee et al. 2015), where we count the number of positive, negative, and neutral words in the input as well as the average of the lexical polarity as four numeric *sentiment features*, using an opinion lexicon provided by the NLTK package (Loper and Bird 2002).

(2) RNN with attention achieves state-of-the-art results on abuse detection on multiple public datasets as reported in (Chakrabarty, Gupta, and Muresan 2019). Although structurally it is similar to our model in Fig. 1, it lacks the novel

	SVM				RNN baseline with attention		RNN with attention supervision (C+S)		
Train data	C		C+S		C	C+S	Encoded loss	L1 loss	L2 loss
Sentiment	Yes	No	Yes	No	No	No	No	No	No
ROC AUC	0.756	0.750	0.782	0.774	0.796	0.803	**0.826**	0.814	0.810
PR AUC	0.581	0.572	0.614	0.606	0.585	0.633	**0.654**	0.638	0.636
F1 score	0.554	0.545	0.544	0.540	0.584	0.615	**0.624**	0.618	0.608

Table 2: Abusive language detection performance. C: using only comment labels, and C+S: using comment and sentence labels.

attention encoder and the attention supervision that we propose in our model.

Preprocessing. We normalized the YouTube comments with a text normalization tool (Özcan 2016), replaced emoticons and urls with special symbols, and finally tokenized and (POS) tagged the comments with the CMU social text tagging tool (Owoputi et al. 2013).

Evaluation metrics. We consider the label "abusive" to be the positive class and use three metrics in our evaluation—the area under the receiver operating characteristic curve (ROC AUC), the area of the precision-recall curve (PR AUC) and the F1 score. The ROC AUC measures the area under the true positive vs. false positive rate curve. The PR AUC measures the area under precision vs. recall curve. PR AUC is known to be a better metric than ROC AUC at comparing algorithms when negative samples (benign comments in our case) are much more than positive samples (abusive comments) (Davis and Goadrich 2006). Unlike F1 score that requires a specific decision-making threshold set on the test data, ROC AUC and PR AUC are free from any threshold tuning.

Detection Results

We train and evaluate all systems on 5 train-test splits to reduce randomness, and report the average performance in Table 2. Since we have both comment-level and sentence-level annotations, we report the performance of SVM and RNN baselines trained on labeled comments alone (denoted as "C" in Table 2) and the performance by using both labeled comments and labeled sentences (denoted as "C+S"). As for the proposed RNN with attention supervision, it is only trained on the labeled comments with access to the labels of the components sentences.

We note that RNNs always outperform the SVM classifier. For the SVM classifier, we find that adding sentiment information does not lead to obvious improvements. By comparing the models trained on C alone, and on C+S, we observe that sentence-level annotations improve P-R AUC for SVM and the RNN baseline.

We observe that attention supervision makes better use of sentence-level annotations given that the RNN with encoded attention loss outperforms the RNN baseline trained on C+S by 2.3% in ROC AUC, by 2.1% in PR AUC and 0.9% in F1 score. The performance gains are statistically significant at p-value of 0.05 using Student's t-test. Moreover, for the model with supervised attention, it is notable that the use of encoded loss is better than both L1 and L2 loss. The gains of the model using encoded attention loss over model instances trained with L1 or L2 are also statistically significant.

Attention Evaluation

We saw how the model trained with attention supervision resulted in improved abuse detection. To provide a comprehensive view of the model's performance, we evaluate the model's ability to learn the correct abusive patterns, which is reflected in segments with high attention.

Qualitative evaluation. We evaluate models' attention on sentence segments. The attention assigned by the RNN model *without* attention supervision is depicted in Fig. 2(a). We compare this with the attention distribution of the model trained with encoded attention loss. As shown in Fig. 2(b), the abusive pattern "you're a cunt" was captured by the model with encoded attention loss. Notably, this example illustrates how our annotations at the sentence-level, also help with phrase-level heterogeneity as well.

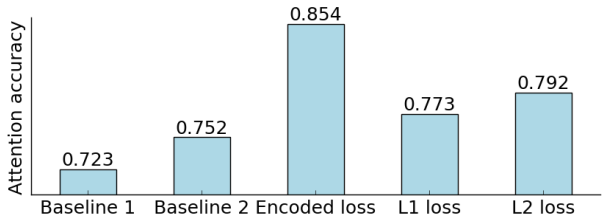


Figure 3: Attention evaluation on test comments.

Quantitative evaluation. Next we quantitatively evaluate the predicted attention over the test comments by analyzing the model's attention weights over the component sentences of the comments. We average the attention weights of the words within each sentence to yield the sentence attention weight. For each abusive comment, we select the sentence with the highest attention weight as the predicted abusive sentence. Then we evaluate the accuracy of the abusive sentence prediction by comparing with the gold labels, yielding the percentage of automatically selected sentences which were manually annotated as abusive.

We report the accuracy of the models with encoded, L1 and L2 loss in Fig. 3, including the two baselines trained without attention supervision—baselines 1 and 2 as the RNN with attention, trained using C and C+S respectively.

We note that baseline 2 captures the abusive patterns more accurately than baseline 1, showing that sentence-level annotation helps abusive segment detection. It is also noteworthy that the model with encoded attention loss outperforms baseline 2 (trained without attention supervision). Even though baseline 2 used both comment- and sentence-level labels, it was trained on isolated sentences without considering the contextual information. This highlights the effectiveness of

	Multi-task					Single-task				
Attention supervision	Encoded	L1	L2	Baseline 1	Baseline 2	Encoded	L1	L2	Baseline 1	Baseline 2
Gender	**0.643**	0.601	0.613	0.585	0.608	0.609	0.599	0.601	0.576	0.582
Race	**0.551**	0.505	0.503	0.483	0.505	0.354	0.323	0.330	0.168	0.307
Appearance	**0.788**	0.760	0.773	0.752	0.760	0.755	0.745	0.737	0.738	0.733
Ideology	**0.610**	0.577	0.559	0.496	0.508	0.511	0.524	0.512	0.477	0.499

Table 3: PR AUC of abuse categorization with and without attention supervision in single- and multi-task settings.

attention supervision for learning the abusive patterns in the context of the entire comment.

Abusive Language Categorization

A fine-grained categorization of abusive comments provides insights into the nature of abusive language. We manually classified the abusive comments into the category set $C = \{\text{gender}, \text{race}, \text{appearance}, \text{ideology}\}$.

Model

Previous work on categorization trained a classifier for each category independently (Van Hee et al. 2015; Dinakar, Reichart, and Lieberman 2011). However, poorly represented categories (e.g., *race* in our data) make training a good classifier for such a category difficult. We adopt the technique of multitask learning, where the main idea is to share information among multiple related tasks so as to improve the model's generalizability of the individual tasks (Standley et al. 2020). In our multitask model, the different categories share information by sharing their lower-level layers (i.e. embeddings in the input layer and the recurrent layer). The predictions for each category are made separately in their respective output layers. We empirically show the resulting performance gain for all categories. Notably, we find that supervised attention helps not only in abuse detection but also in categorization.

The model for categorization was similar to that used for abuse detection (Fig. 1), except the single two-dimensional output vector $\mathbf{y}'$ was replaced with four two-dimensional vectors $\{\mathbf{y}'_\mathbf{c}\}_{\mathbf{c}\in\mathbf{C}}$, each two-dimensional vector $\mathbf{y}'_\mathbf{c}$ corresponding to category c. We used cross-entropy loss as category c's prediction loss L_c. The total loss was again the sum of the prediction loss and the attention loss:

$$L = \sum_{c\in C} \omega_c L_c + \beta L_a, \tag{7}$$

where ω_c is the weight of category c, and $\sum \omega_c = 1$. In multitasking, there is a primary category c with a higher weight ω_c than the weights $\omega_{c'}$ for the auxiliary categories c'. We report the per-category performance by taking each category as the primary category respectively. The hyperparameters were tuned on the validation data, with $\beta = 0.2$, $\omega_c = 0.7$, and $\omega_{c'} = 0.1, \forall c' \neq c$.

Experiments

As before, for our experiments on *categorizing* abusive language, we used a standard RNN model with attention as a strong baseline. Baseline 1 was trained on C, and baseline 2 was trained on C+S. A third model is an RNN model with the same idea of attention supervision (used for the classification task) but now in a multitask learning set-up described above.

We evaluated the models with 5 train-test splits, and report their average performance in Table 3. All the systems were RNNs with different attention losses in either a single-task or a multi-task setting. We report the PR AUC of each category for each system, and evaluate how supervised attention and multitask learning affect the performance. Overall, baseline 2 achieves better PR AUC than baseline 1 due to the extra sentence-level annotations. Attention supervision with encoded loss makes better use of sentence annotations than systems with other attention losses as well as the baselines without attention loss.

Comparing the models with and without attention supervision, we note that attention supervision improves categorization in both single- and multi-tasking scenarios (all are absolute gains); the highest improvement was seen in the poorly represented categories of *race* and *ideology*. For the *race* category, the supervision with encoded loss improves the PR AUC by 4.7% over baseline 2 in single tasking, and 4.6% in multitasking. As for *ideology*, the encoded attention loss yields a gain of 10.2% over baseline 2 in multitasking.

Multi-task learning improves categorization in all categories; we see an increase of 19.7% in the performance of the race category when encoded attention loss is applied, an increase of 31.5% in baseline 1, and an increase of 19.8% in baseline 2. Note that all gains reported are absolute.

The best-performing system is the combination of encoded attention loss with multi-task learning. It uses essentially the same training data as baseline 2. Compared with baseline 2 without attention supervision in single tasking, it increases the PR AUC by 6.1% in the gender category, 24.4% in race, 5.5% in appearance, and 11.1% in ideology.

Conclusion and Limitations

We have presented a new annotated dataset of abusive language from YouTube, as well as an empirical study on the use of supervised attention of neural networks to improve the detection and categorization of abusive language.

A primary limitation of our methodology is that our data comes only from feminism-related channels, which introduced bias and limits the generality of our results. Moreover, due to limitations of the annotation interface, the thread structure was not available to annotators, and they did not follow links in the comments or view the associated videos. This was intentional, so that the automatic detection would be based solely on textual information. Hence, two important directions for future work are to (a) study the performance of supervised attention on a broader class of datasets, and (b) conduct a joint analysis of text *and* the accompanying media.

Acknowledgements

This work was supported in part by the National Science Foundation under grant no. 1720268. We would like to thank our annotators: Cagil Torgal, Ally Montesino, Lauren Fisher, Kaylie Skinner, Lital Hartzy, Madison Kohler, Gabriele Mamone, Abigail Matterson, Hannah Phillips, Palmer Tirrell, Victoria Wiliams, Kristina Youngson, Huibin Zhang and Talia Akerman, and our participants: Luz Robinson, America El Sheikh, Uma Kumar, Savannah Herrington, Briana de Cola, Ciara Tobin, Angela Rodriguez, Carmen Florez, Sky Martin, Caroline Spitz, Claudia Rodriguez and Paige Hespe. We would also like to thank Sreedhar Radhakrishnan and Ganesh Ramadurai for their help in ensuring the reproducibility of our results and comparison against data from Stormfront.

References

Alakrot, A.; Murray, L.; and Nikolov, N. S. 2018. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science* 142: 174–181.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760.

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1903–1913.

Chakrabarty, T.; Gupta, K.; and Muresan, S. 2019. Pay "Attention" to your Context when Classifying Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, 70–79.

Choi, E.; Johnson, D. A.; Moon, K.; and Oah, S. 2018. Effects of positive and negative feedback sequence on work performance and emotional responses. *Journal of Organizational Behavior Management* 38(2-3): 97–115.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. CONAN-COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240. ACM.

de Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 11–20.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11(02): 11–17.

Ferguson, M.; and Barry, B. 2011. I know what you did: The effects of interpersonal deviance on bystanders. *Journal of Occupational Health Psychology* 16(1): 80.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)* 51(4): 85.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; Gnanasekaran, R. K.; Gunasekaran, R. R.; et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, 229–233.

Hoff, D. L.; and Mitchell, S. N. 2009. Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration* 47(5).

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

James, D.; Flynn, A.; Lawlor, M.; Courtney, P.; Murphy, N.; and Henry, B. 2011. A friend in deed? Can adolescent girls be taught to understand relational bullying? *Child Abuse Review* 20(6): 439–454.

Joksimovic, S.; Baker, R. S.; Ocumpaugh, J.; Andres, J. M. L.; Tot, I.; Wang, E. Y.; and Dawson, S. 2019. Automated Identification of Verbally Abusive Behaviors in Online Discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, 36–45.

Kowalski, R. M.; and Limber, S. P. 2007. Electronic bullying among middle school students. *Journal of adolescent health* 41(6): S22–S30.

Liu, L.; Utiyama, M.; Finch, A.; and Sumita, E. 2016. Neural Machine Translation with Supervised Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3093–3102.

Liu, S.; Chen, Y.; Liu, K.; and Zhao, J. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1789–1798.

Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Low, K. D.; Radhakrishnan, P.; Schneider, K. T.; and Rounds, J. 2007. The experiences of bystanders of workplace ethnic harassment. *Journal of Applied Social Psychology* 37(10): 2261–2297.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421.

Maity, S. K.; Chakraborty, A.; Goyal, P.; and Mukherjee, A. 2018. Opinion conflicts: An effective route to detect incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–27.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2): 1–25.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, 145–153.

Nockleby, J. T. 2000. Hate speech. *Encyclopedia of the American constitution* 3(2): 1277–1279.

Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, 380–390.

Özcan, S. 2016. Tweet-preprocessor. http://preprocessor.readthedocs.org/. Accessed: 2020-08-01.

Park, J. H.; and Fung, P. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, 41–45.

Patchin, J. W.; and Hinduja, S. 2015. Measuring cyberbullying: Implications for research. *Aggression and Violent Behavior* 23: 69–74.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 1125–1135.

Peter, I.-K.; and Petermann, F. 2018. Cyberbullying: A concept analysis of defining attributes and additional influencing factors. *Computers in human behavior* 86: 350–366.

Price, M.; Dalgleish, J.; et al. 2010. Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. *Youth Studies Australia* 29(2): 51.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. 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)*, 4757–4766.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021. The Evolution of the Manosphere Across the Web. In *Proceedings of the 15th International AAAI Conference on Web and Social Media*.

Schmidt, A.; and Wiegand, M. 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*, 1–10.

Sennrich, R.; and Haddow, B. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, volume 1, 83–91.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, 9120–9132. PMLR.

Van Hee, C.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; and Hoste, V. 2015. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, 672–680.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, 80. Association for Computational Linguistics.

Wang, B. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2311–2320.

Waseem, Z. 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*, 138–142.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yin, D.; Xue, Z.; Hong, L.; Davison, B. D.; Kontostathis, A.; and Edwards, L. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2: 1–7.