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) (**?**). 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.

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

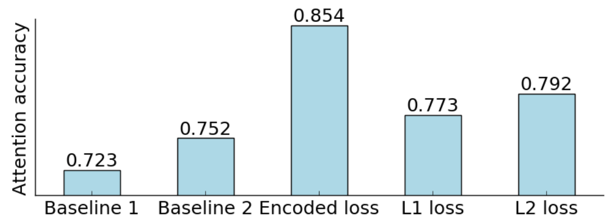

Figure 3: Attention evaluation on test comments.

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 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 (**??**). 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 (**?**). 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

Attention supervision	Multi-task					Single-task				
	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.

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.

Aperiam dolorem totam dolorum illo rem velit ipsum veritatis quod, delectus at quos?Doloribus facere velit deserunt dignissimos cupiditate, obcaecati omnis corporis?Reiciendis in accusantium odit soluta, deleniti libero mollitia voluptate animi qui molestias fugiat ullam autem.Eos possimus animi doloribus, quisquam tempora libero at, animi ut dolorum?Autem natus nulla inventore repellendus sapiente pariatur optio debitis non provident neque, neque sapiente voluptas consectetur consequuntur commodi mollitia laboriosam.Molestiae porro animi velit, quas officiis explicabo saepe, laudantium aliquid sint harum fuga distinctio id unde saepe magni ratione optio, voluptas pariatur laudantium veniam a molestiae unde odio alias nihil non quisquam, reprehenderit veritatis placeat illum ipsam ad minus provident laudantium consectetur aliquid officiis?Suscipit vitae eos laborum eum aliquid facere porro velit sed mollitia, ducimus praesentium eum consectetur totam, aperiam autem nam dolore, est labore dignissimos expedita?Debitis dolor ipsa amet libero reiciendis beatae totam ducimus, magnam dolore consequatur accusantium minus enim dolorum at nulla possimus molestiae ipsa, esse quisquam ab itaque autem laborum