Empirical Analysis of Multi-Task Learning for Reducing Identity Bias in Toxic Comment Detection

Ameya Vaidya,[1] **Feng Mai,**[2] **Yue Ning**[3]
[1]Bridgewater-Raritan Regional High School
[2]School of Business, Stevens Institute of Technology
[3]Department of Computer Science, Stevens Institute of Technology
ameyav993@gmail.com, {fmai, yue.ning}@stevens.edu

Abstract

With the recent rise of toxicity in online conversations on social media platforms, using modern machine learning algorithms for toxic comment detection has become a central focus of many online applications. Researchers and companies have developed a variety of models to identify toxicity in online conversations, reviews, or comments with mixed successes. However, many existing approaches have learned to incorrectly associate non-toxic comments that have certain trigger-words (e.g. gay, lesbian, black, muslim) as a potential source of toxicity. In this paper, we evaluate several state-of-the-art models with the specific focus of reducing model bias towards these commonly-attacked identity groups. We propose a multi-task learning model with an attention layer that jointly learns to predict the toxicity of a comment as well as the identities present in the comments in order to reduce this bias. We then compare our model to an array of shallow and deep-learning models using metrics designed especially to test for unintended model bias within these identity groups.

Introduction

The identification of potential toxicity within online conversations has always been a significant task for current platform providers. Toxic comments have the unfortunate effect of causing users to leave a discussion or give up sharing their perspective and can give a bad reputation to platforms where these discussions take place. Twitter's CEO reaffirmed that Twitter is still being overrun by spam, abuse, and misinformation (Stelter 2018). To deal with this problem, researchers and companies have done extensive investigations into the field of toxic comment detection. Current research involves tackling common challenges in toxic comment classification (van Aken et al. 2018), identifying subtle forms of toxicity (Noever 2018), detecting early signs of toxicity (Zhang et al. 2018), and analysing sarcasm within conversations (Ghosh, Fabbri, and Muresan 2018).

Over the past few years, a variety of models and methods have been proposed to detect online toxic comments. Current baseline methods exploit the representation of documents as character n-grams or TF-IDF (Badjatiya et al. 2017) which are then learned by Logistic Regression or Support Vector Machines (Noever 2018). Recently, deep learning methods such as convolutional neural networks (Georgakopoulos et al. 2018) and recurrent neural networks (Zhou et al. 2016) have been popularized in natural language processing to analyze online content. Furthermore, bidirectionality (Zhou et al. 2016), attention mechanisms (Bahdanau, Cho, and Bengio 2015), and ensemble learning (Dietterich 2000) have also shown improved performance in text sentiment analysis.

Table 1: Example of toxic comments with identity attack where Identity can be replaced by "gay", "black" etc.

Identity	Toxic(+)	Non-toxic(-)
Positive	*⟨ Identity ⟩ people are gross and universally hated!*	*I am a ⟨ Identity ⟩ person, ask me anything.*
Negative	*What the heck is wrong with you?*	*Thanks for the help. I really appreciate it!*

However, many existing works have documented that current toxic comment classification models introduce bias into their predictions. They tend to classify comments that reference certain commonly-attacked identities (e.g., gay, black, muslim) as toxic without the comment having any intention of being toxic (Dixon et al. 2018; Borkan et al. 2019b) as shown in Table 1. For example, the comment "I am a black woman, how can I help?" might be classified by a model as toxic because it references the 'black' identity. Furthermore, the Conversation AI team at Google Jigsaw has acknowledged that their Perspective API framework, which attempts to detect toxicity in online conversations, seems to generate higher toxicity scores for sentences containing commonly targeted identity groups.[1] Current research efforts to investigate model bias (Sap et al. 2019; Davidson, Bhattacharya, and Weber 2019) have detected a correlation between race, identity, and model predictions in the context of hate and abusive speech. Evaluation metrics

[1] https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

have also been developed to test specifically for implicit model bias (Dixon et al. 2018). However, not many novel mitigation solutions have been proposed within current research efforts, which is something that we hope to contribute with the proposal of our model.

In this paper, our main focus is to reduce the false positive rate on non-toxic comments that make reference to identities known historically and empirically to introduce model bias. Our empirical analysis focuses specifically on the improvement of the following identities because of their tendency to be associated with a high false positive rate: gay, lesbian, bisexual, transgender, black, muslim, and jewish. To deal with this challenge, we propose a multi-task learning framework that simultaneously identifies toxicity and identity information within a comment. Learning these tasks jointly will allow the model to share common patterns and better distinguish between toxic and non-toxic comments that reference these identities. This paper also aims to evaluate various shallow and deep learning models adapted from existing literature, including logistic regression and recurrent neural networks, on the task of mitigating bias. We evaluate the proposed multi-task learning model and other deep learning methods on a dataset of 1,804,874 unique comments published by Google Jigsaw during Kaggle's Unintended Bias in Toxicity Classification Challenge.[2] To keep our focus on mitigating unintended bias, we utilize a set of evaluation metrics that are specifically designed for measuring bias in the model outputs.

The ultimate goal of our research is to help maintain the civility of conversations on common social media platforms while minimizing the amount of non-toxic comments that are classified as toxic. Our main contributions are summarized as below:

1. We perform an empirical study for a multitude of classifiers on a new public dataset containing over 1.8 million comments. We also compare classifiers with the specific focus of reducing unintended model bias within online conversations.
2. We propose a multi-task learning model that outperforms other models at mitigating unintended bias, especially for certain identities that historically bring a high rate of false positives. The attention layer included in the multi-task learning model allows us to capture hidden state dependencies and distinguish between toxic and non-toxic comments. We also employ a custom-weighted loss function that allows our model to increase penalization on false positive mistakes.
3. We analyse the classifiers' predictions on a variety of evaluation metrics. These measures include F1-measures and the AUC-ROC score. In addition, we evaluate our models on metrics designed specifically for unintended model bias: Generalized Mean Bias AUC, Subgroup AUC, and BPSN AUC. We also compare non-toxic and toxic comments across models and with Google's Perspective API.

In the following sections, we will introduce the related work and the investigated dataset followed by the proposed multi-task learning framework. Then we discuss the experimental evaluation and results. Finally we conclude our work with future directions.

Related Work

In this section, we will briefly review recent developments in multi-task learning. We will then focus on new attempts on toxic comment classification and identity bias in natural language processing models.

Multi-Task Learning

Multi-task learning (Caruana 1998; Argyriou, Evgeniou, and Pontil 2007) has been widely studied and applied in natural language processing (NLP) (Collobert and Weston 2008; Deng, Hinton, and Kingsbury 2013), computer vision, and other machine learning applications (Ramsundar et al. 2015). In deep learning models, multi-task learning is usually implemented by either sharing hidden layer model parameters (Long et al. 2017) or regularizing parameters among related tasks to be similar (Duong et al. 2015). Recent works show that multi-task learning can improve performance on various NLP tasks while revealing novel insights about language modeling (Søgaard and Goldberg 2016). In terms of network architecture, our work is closest to the LSTM-based multi-task learning frameworks (Liang and Shu 2017; Suresh, Gong, and Guttag 2018). However it is known that the performance of multi-task learning is task specific (Misra et al. 2016). Which framework is more effective at teasing out identity information while detecting toxic comments is an open empirical question.

Toxic Comment Detection

Machine learning for detecting toxic comments has been a significant focus in Natural Language Processing research over the past few years. This is in part due to the availability of large corpora of online social interactions. Wikimedia Foundation (Wulczyn, Thain, and Dixon 2017) released an annotated dataset of personal attacks, toxic messages, and aggression from the English Wikipedia Talk pages. Google Jigsaw also published two Kaggle competitions which have allowed researchers to gain access to datasets with 2.5 million training examples of toxic comments. In terms of methods, most research takes a text classification approach similar to sentiment analysis and spam detection (Mishra et al. 2018; Davidson et al. 2017; Wulczyn, Thain, and Dixon 2017). These methods rely on document features (readability, emotion, sentiment, n-grams), author features (demographics, social network positions), or contextual features (the relationship of a document to others) to train classifiers.

More recent research advances toxic comment detection models on two fronts. Some studies move beyond using documents as the units of analyses and model the behavior of the users (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015), or take a more proactive approach to detect online conversations that are susceptible to escalation (Zhang et al.

[2] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview

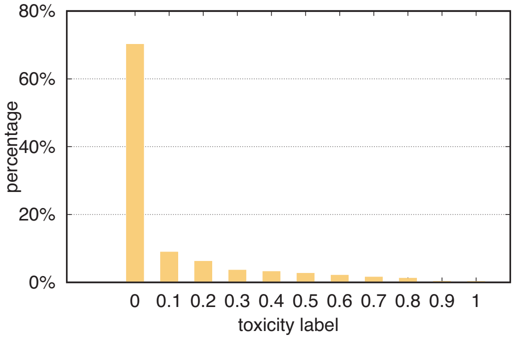

Figure 1: Percentage distribution of toxicity labels in the dataset. X-axis is the interval of toxicity scores (e.g. '0'=[0,0.1)) and Y-axis is the combined percentages of comments in each interval. It shows a clear imbalanced distribution of toxic and non-toxic comments in the dataset.

2018). Another stream of work uses neural network models to classify toxic comments and has shown impressive results (Georgakopoulos et al. 2018; Chen, McKeever, and Delany 2019; Elnaggar et al. 2018; Srivastava, Khurana, and Tewari 2018). Although these new models can achieve good performance without hand-crafted features, a potential downside is that the decisions made by the classifiers are more opaque. When the model is deployed, it may conflate identity attacks with identity disclosures, and make a biased decisions against the latter. Our work extends this stream of research and uses multi-task learning to explicitly account for the identity bias.

Unintended Identity Bias in NLP Models

A growing number of studies have called attention to the identity related biases in natural language models. Several studies have highlighted how word embeddings exhibit human stereotypes towards genders and ethnic groups (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). One way to counter such biases is to enhance the interpretability of black box models (Guidotti et al. 2019) so that humans can intervene when a model makes an unfair decision. Another way to address the issue, which is the focus of this study, is to design models to circumvent protected identity attributes. Several methods have been proposed in the context of structured or numerical data (Corbett-Davies and Goel 2018), but methods applicable to text data generated by online users are rare. During our research, we discovered the Pinned AUC metric which is popularly used to mitigate unintended bias (Dixon et al. 2018). However, we have decided its use is unwarranted in this paper due to recent discoveries which suggest that Pinned AUC can be distorted by uneven distributions, which is prevalent within the dataset we analyze (Borkan et al. 2019a).

Dataset

We analyse a dataset published by the Jigsaw Unintended Bias in Toxicity Classification Challenge on Kaggle. It contains 1,804,874 comments annotated by the Civil Comments platform. Each comment is shown up to 10 annotators who classify each comment as either *Very Toxic*, *Toxic*, *Hard To Say*, or *Not Toxic*. Each comment is then given a toxicity label based on the fraction of annotators that classified it as either *Toxic* or *Very Toxic*. For evaluation, every comment with a toxicity label greater than or equal to 0.5 was considered to be toxic (the positive class). As discussed by Jigsaw, a toxic comment usually contains rude, disrespectful, or unreasonable content that is somewhat likely to make you leave a discussion or give up on sharing your perspective.

Table 2: Number of comments labeled with each identity and percent of comments in each identity that are non-toxic.

Identities	**Count**	**Non-Toxic**
male	64,544	90.40%
female	55,048	90.86%
homosexual (gay or lesbian)	11,060	80.99%
christian	40,697	94.41%
muslim	21,323	85.06%
jewish	7,669	89.75%
black	17,161	80.31%
white	28,831	82.24%
psychiatric or mental illness	6,218	85.91%
All Identities	191,671	85.56%

In addition, each comment was also labeled with a multitude of identity attributes (non-exclusive), which demonstrates the presence of a specific identity in a comment. These identities include *male*, *female*, *homosexual (gay or lesbian)*, *christian*, *jewish*, *muslim*, *black*, *white*, and *psychiatric or mental illness*. Label values were given based on the fraction of annotators who believed a comment fit the identity mentioned. Each comment was also labeled with five subtype attributes: *severe_toxicity*, *obscene*, *threat*, *identity_attack*, and *insult* based on the percent of annotators who identified a comment with the aformentioned subtype. Out of these nine identities, we found that the following identities tended to have the highest false positive rates: *homosexual*, *muslim*, *jewish*, *black*.

Figure 1 suggests that the distribution of the toxicity label within the dataset follows a long tail distribution. Approximately 92% of the comments are classified as non-toxic (negative class). Table 2 shows the distribution of identity labels in the dataset. In addition, within this dataset, the average document length is approximately 51.28 words long, meaning that identifying long-range dependencies is an important consideration in this paper. We discuss the use of Long Short-Term Memory Networks (LSTM) later in this paper to deal with this challenge.

Models and Tasks

In this section, we explore and propose a multi-task learning framework whose focus is to improve the accuracy of correctly detecting toxic comments by jointly learning toxicity and identity information. The toxicity task aims to correctly predict the toxicity score for a comment. The identity task is designed to predict the presence of an identity in a comment.

These tasks work jointly to reduce the model bias towards commonly attacked identities in Table 2.

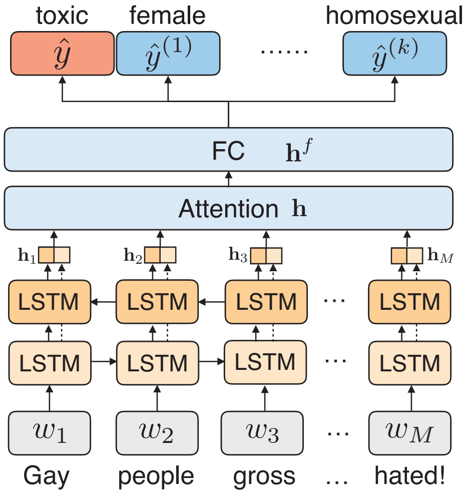

Figure 2: An overview of the proposed MTL frameworks. 'FC' indicates a fully connected layer. Each label on the top represents a task.

Model The overview of the model is illustrated in Figure 2. The highlights of our model include an embedding layer, two Long Short-Term Memory Network (LSTM) layers, an attention mechanism, and a custom loss function. Our multi-task learning model utilizes deep sharing to jointly learn the toxicity and identity tasks which allows us to exploit the commonalities and differences between tasks. The embedding layer allows our model to gain a better understanding of the semantics encoded within each word. The LSTM layers and attention mechanism work together to capture long-range and hidden-state dependencies, form a complete understanding of the entire document by parsing individual words, and pay specific attention to words that are relevant to each task. Finally, the custom-weighted loss function allows our model to place extra focus on learning to mitigate unintended bias, rather than simply increasing the ROC-AUC score for the entire test set. We are one of the first to implement a multi-task learning model that makes use of all of these components specifically to investigate the mitigation of unintended bias. Each of these elements is explored in more detail below.

Embedding & LSTM layers Each word in a sentence is converted to a word embedding vector of dimension D which concatenates two parts: 1) pre-generated embeddings from the global vectors for word representation (Pennington, Socher, and Manning 2014) and 2) pre-generated embeddings from the vectors provided by FastText (Joulin et al. 2016). Assuming there are N total comments in the training dataset, each comment example has M words (M = max length) and is represented as $\mathbf{s} = [\mathbf{x}_1, ..., \mathbf{x}_M]$. Each comment is associated with a toxicity label y and a set of identity labels $y^1, ..., y^K$ (K = number of identity labels). Each word $\mathbf{x}_m \in \mathbb{R}^D$ is represented by an embedding vector. Then we apply a bi-directional recurrent neural network (e.g. LSTM), a forward LSTM and a backward LSTM, to the sentence $\mathbf{s}$. We obtain the hidden state $\mathbf{h}_m$ for each word $\mathbf{x}_m$ by concatenating the forward hidden state $\overrightarrow{\mathbf{h}}_m$ and the backward hidden state $\overleftarrow{\mathbf{h}}_m$.

Attention Attention mechanisms have shown to produce state-of-the-art results in many natural language processing tasks such as machine translation (Bahdanau, Cho, and Bengio 2015) when combined with neural word embeddings. In this paper, we explore a feed-forward attention mechanism (Raffel and Ellis 2016) on the bidirectional LSTM to "memorize" the influence of each hidden state:

$$a_m = \frac{\exp(\tanh(\mathbf{W}_a \mathbf{h}_m))}{\sum_{j=1}^{M} \exp(\tanh(\mathbf{W}_a \mathbf{h}_j))}, \tag{1}$$

$$\mathbf{h} = \sum_m a_m \mathbf{h}_m, \tag{2}$$

where a_m measures the importance of current word m and $\mathbf{W}_a$ is the weight parameter to be learned. Then two fully connected dense layers are applied on the hidden state of the sentence (comment) $\mathbf{h}$. This attention mechanism will allow our model to directly access the entire sequence. Our model will pay more "attention" to the words that correlate with toxicity and identity labels. Thus it helps to avoid classifying these comments without a deeper and more meaningful understanding of each document.

Multi-Task Learning Rather than learning each task individually, learning multiple tasks simultaneously has been theoretically and empirically proven to improve prediction performance (Caruana 1993). Multi-task learning works the best when multiple tasks are related in some shape or form (Argyriou, Evgeniou, and Pontil 2007). In order to reduce unintended model bias, we take advantage of multi-task learning to model related tasks and capture their internal patterns. For instance, when predicting the toxic comment "gay people are gross and universally hated", the toxicity task will focus on the toxic elements "gross and universally hated" while the identity task will identify the trigger word "gay" in the comment. We expect involving identity tasks will reduce model bias by mitigating the confusion between identity and toxicity in predictions. Our model will utilize hard-parameter sharing as specified in (Ruder 2017) which prevents the model from overfitting and allows for more opportunities to share information between the toxicity and identity classifications.

Prediction As shown in Figure 2, for different tasks, we share the same structure of the network till the last output layer. The predictions for toxic label $\hat{y}$ and identity labels $\hat{y}^k (k = 1...K)$ are then modeled as below:

$$\hat{y} = \sigma(\mathbf{w}_t^\top \mathbf{h}^f + b_t), \tag{3}$$

$$\hat{y}^k = \sigma(\mathbf{w}_k^\top \mathbf{h}^f + b_k), k = 1...K, \quad (4)$$

where $\mathbf{h}^f$ is the output from the dense layers. We have also evaluated different levels of sharing. For instance, another design is sharing the bi-directional LSTM layers among tasks but each task has its own fully connected layers. We observe that the best performance comes from the deep sharing mechanism as demonstrated as Eq. 3 and 4.

Model loss Finally, we design a weighted binary cross-entropy (CE) loss for all the label estimates of training examples. Given that a comment can have multiple identity labels, a general cross-entropy is not used in this case:

$$L = \sum_{n=1}^{N} \beta_n \big[\alpha J_{\text{CE}}(\hat{y}_n, y_n) + (1-\alpha) \sum_{k=1}^{K} J_{\text{CE}}(\hat{y}_n^k, y_n^k)\big]. \quad (5)$$

We employ a weighted loss per example (β_n) and per task (α). By default, $\beta_n = 1$. If an example is a non-toxic example with identity information, its weight $\beta_n = \beta_n \times c$ where c is a constant (e.g. $c = 3$ in our experiments). The task weight $\alpha \in [0, 1]$ is selected by a grid search in validation set. All model parameters are trained via back-propagation and optimized by the Adam algorithm (Kingma and Ba 2015) given its efficiency and ability to avoid overfitting.

Experimental Study

The purpose of our experiment is to compare the performance of our multi-task learning model to other baseline models. The four types of baseline models used for comparison are: Logistic Regression, CNNs, LSTMs, and GRUs. Our hypothesis is that the multi-task learning model will be able to outperform the other baseline models in multiple categories, especially in those that measure unintended bias. In this experiment we focus on the toxicity task and the toxicity scores predicted by the model rather than the identity scores.

Experiment Setup

Text preprocessing. Before the model training, we perform some basic preprocessing on the data. To convert the raw text to a usable format, we first tokenize the comments. Because comments vary in length, the max-length is defined as 220 words. Sequences that had less than 220 words are padded with zeroes. During the process of tokenization, each comment is stripped of certain punctuation marks but was not converted to lowercase.

Model-specific preprocessing. We also perform preprocessing specific to the multi-task learning model. Because only 405,130 out of 1,804,874 comments are annotated for each of the identities, we need to fill in the scores for the rest of the identities in order to employ an effective multi-task model. To fill in the rest of the identities, we train a multi-class classifier on the ˜400,000 training examples with the annotated identity scores to predict the identity scores for the remaining ˜1.4 million training examples. This multi-task model employs the same architecture as the MTL model that we discussed in the last section. However, we omit an attention layer from this model because we found that an attention layer does not significantly improve the accuracy of predicting the identities within a comment. The results from our MTL model is then fed into our multi-task learning model for prediction as shown in Figure 3.

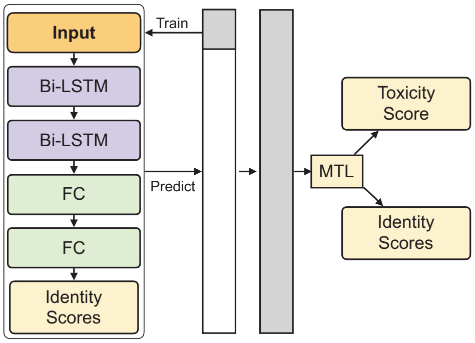

Figure 3: Multi-task learning model with model-specific preprocessing for propagating identity labels.

Cross-Validation In our experiments, we perform K-fold (K=5) cross-validation on the dataset. In each fold, 80% of the data is set aside for training and 20% is used for validation, which translates into roughly 1.4 million and 400,000 comments respectively.

Model parameters and hyper-parameter settings We select α (model loss eq.) by a grid search in our validation set and α is chosen to be 0.6 with the best performance. We set the dimension of hidden states in the two bidirectional LSTM layers to be 256 and 512 for the two fully-connected layers. Rectified activation functions are applied to the fully-connected layers and a sigmoid activation function is applied to the output layers. We also introduce a spatial dropout of 20% between the embedding and first bidirectional layer.

Comparison Methods (Baseline Models)

Each of the following baseline models was developed using the Keras framework. A significant portion of each of the following baseline models was adapted from existing works, albeit small changes in number and size of layers used.

- **Logistic Regression** Logistic Regression (Neter et al. 1996) has widely been used for binary classification tasks. For text classification tasks, documents are usually vectorized into bag-of-words (BoW) features (e.g. TF-IDF). As a comparison to dense vectors in deep learning models, our model applies a TF-IDF vectorizer to the raw comments and then passes it through a standard logistic regression model to obtain the final predictions.
- **Convolutional Neural Networks** Convolutional Neural Networks (LeCun et al. 1998) have proved to be very successful when it comes to sentence or character-level sentence classification (Kim 2014). CNNs have been known

to work better for datasets with a large amount of training examples and can work well for user-generated data, given its ability to deal with the "obfuscation of words" in comments and "detect specific combinations of features" in text classification. Our CNN model is adapted from the following paper (Georgakopoulos et al. 2018).

- **Long Short-Term Memory Network** LSTMs (Hochreiter and Schmidhuber 1997) were introduced primarily to overcome the problem of the vanishing gradient. As a variant of Recurrent Neural Networks (RNN), it has proven to have a better ability to learn long-range dependencies. In the Simple LSTM baseline model, we introduced a 20% spatial dropout. The input is passed through two LSTM layers of 256 units each. Afterwards, the input passes through two dense layers of 512 units each with a rectified linear activation function. Finally, we obtain a single output (toxicity score) by applying a sigmoid activation function to the final dense layer. The architecture for LSTMs and GRUs were adapted from the following paper (van Aken et al. 2018).
- **Gated Recurrent Unit** GRU (Chung et al. 2014) operates similarly to an LSTM but instead uses a reset and update gate, where the reset gate acts to forget the previous state and the update gate decides how much of the candidate activation to use in updating the cell state. Our GRU model is similar to the structure of our LSTM model, with the exception that only 128 units were used per GRU layer.
- **Bidirectionality** Introducing bidirectionality into a RNN can help a network learn from both past and future context (Schuster and Paliwal 1997). In this architecture, two layers of hidden nodes are introduced. In the second layer, the input is reversed and the sequence is passed backwards into the network. Within the scope of this task, understanding and learning from a sequence in both directions can lead to a more complex and more accurate understanding of the document. In this paper, we implement a Bidirectional LSTM and a Bidirectional GRU. They follow the same structures as the Simple LSTM and GRU specified in the previous paragraphs.
- **MTL-Aux** In addition to the baseline models, we developed another multi-task learning model for comparison. Instead of using the nine identity labels, MTL-Aux focuses on five alternate subtype toxicity attributes: *severe toxicity*, *obscene*, *threat*, *insult*, and *identity attack*. The model follows the exact same structure as MTL-attn but predicts the aforementioned five subtype outputs rather than the nine identity outputs.

Evaluation Metrics

For basic toxicity detection evaluation, we calculate the ROC-AUC, precision, recall, and F1-scores for the full test set. Predictions with toxicity scores that are greater than or equal to 0.5 are considered to be part of the positive (toxic) class and vice-versa.

Unintended bias evaluation metrics are introduced and specified by the Google Conversation AI Team in their paper (Borkan et al. 2019b). The Generalized Mean of Bias AUCs metric was introduced by their Kaggle competition[2]. The following evaluation metrics are specifically crafted to accurately measure the reduction of unintended bias by a model. By restricting the test set, we get a better understanding of how each of the models perform within the scope of toxic comment detection and bias reduction.

- **Subgroup AUC** We restrict the test set to comments for which each identity label is positive. An ROC-AUC score is then calculated for each of the identity groups which is hereby called the Subgroup AUC. A low value indicates that the model does a bad job of distinguishing between toxic and non-toxic comments in the context of that specific identity (e.g. gay, muslim, black).
- **BPSN (Background Positive, Subgroup Negative) AUC** To calculate this metric, we restrict the test set to non-toxic comments that mention the identity and toxic comments that don't mention the identity. We obtain the BPSN AUC by getting the ROC-AUC score from this restricted test set. The main purpose of this metric is to measure the false positive rate of each model in the context of each specific identity. A higher BPSN AUC score means that a model is less likely to confuse non-toxic examples that mention the identity with toxic examples that do not, meaning that the model is able to mitigate biases towards a specific identity. BPSN AUC can be considered a stronger evaluation metric than Subgroup AUC because it tailors towards the specific focus of this paper: reduce the false positive rate towards certain identities.
- **Generalized Mean of Bias AUCs** One overall measure is calculated from the Subgroup AUCs using the following formula: $M_p(m_s) = (\frac{1}{N}\sum_{s=1}^{N} m_s^p)$ where M_p is the p-th power-mean function, m_s is the bias metric calculated for subgroup s, and $N = 9$ which is the number of identity subgroups. We set $p = -5$ as suggested in the competition. A low value indicates model bias toward one or more of the identities. This metric is essentially a average of all nine subgroup AUCs. This metric is hereby referred to as Generalized Mean AUC throughout the rest of the paper.

Results & Empirical Analysis

Proposed Model: MTL-attention

The overall binary classification performance is summarized in Table 3. The MTL-attention model outperforms all other baseline models in Generalized Mean Bias AUC, suggesting that it was the most successful at accurately classifying comments with any of the aforementioned identities present. MTL-attention also outperformed all other models in recall, precision, and F1-score. This strongly suggests that learning tasks in parallel could be useful in forming a shared representation of the dataset, where what is learned for one task helps other tasks to be learned better. In addition, the inclusion of the attention layer and custom loss function enabled the model to pay closer attention to certain words and phrases that signaled the possibility of a false positive classification.

Our results are statistically significant when compared to the best baseline model. We conduct a Kolmorogov-Smirnov

Table 3: Binary classification performance of different methods on toxic comments. Bold face indicates the best result of each column and underlined the second-best. Codes are used in subsequent tables to refer to each of the models.

Model	Generalized Mean AUC	AUC	Precision	Recall	F1-Score	Code
Logistic Regression	0.8999	0.9488	0.79	0.50	0.61	LR
CNN	0.9212	0.9635	0.86	0.47	0.60	CNN
Simple LSTM	0.9267	0.9662	0.85	0.50	0.63	LSTM
Bidirectional LSTM	0.9316	0.9694	0.83	0.55	0.66	Bi-LSTM
Bidirectional LSTM & Attn	0.9322	0.9696	0.84	0.55	0.66	Bi-LSTM-A
Simple GRU	0.9284	0.9676	0.83	0.54	0.65	GRU
Bidirectional GRU	0.9319	0.9637	0.84	0.52	0.64	Bi-GRU
Bidirectional GRU & Attn	0.9325	0.9697	0.84	0.53	0.66	Bi-GRU-A
MTL-Aux	0.9317	0.9693	0.86	0.53	0.65	MTL-Aux
MTL-attention	**0.9407***	**0.9709**	**0.88***	**0.59***	**0.71***	MTL-attn

* Identifies statistical significance ($p < 0.05$) compared to best baseline model in the category.

test to evaluate the difference in means for non-Gaussian results. For Generalized Mean AUC, precision, recall, and F1-score, we found the p-value to be below 0.05 which implies that the improvement in performance between MTL-attention and the best performing baseline model was statistically significant for these evaluation metrics. From these results, we conclude that our multi-task learning model introduces a low level of variance and is effective at reducing unintended model bias when compared to current state-of-the-art models.

While an improvement of 0.8% in the Generalized Mean AUC may seem incremental, it is in fact significant when observing its effect on the rate of false positives. With approximately 191,671 comments being labeled as being associated with an identity and about 14.44% being non-toxic, we have 27,677 non-toxic comments that are associated with an identity. If the main improvement achieved by the multi-task learning model was in reducing bias and the false positive rate (evidenced by an increase in precision), then a 0.8% improvement in Generalized Mean AUC can imply a significant improvement for non-toxic comments. This idea is further explored within the following Subgroup & BPSN AUC subsection.

Baseline Models

The two best baseline models are Bi-LSTM-A and Bi-GRU-A with the highest Generalized Mean AUCs and the highest F1-scores. Bi-LSTM and Bi-GRU follow close behind, suggesting that an attention layer by itself does not contribute significantly to an improvement in model performance. MTL-Aux had comparable performance to Bi-LSTM, suggesting that multi-task learning with the five subtype attributes (*severe toxicity*, *obscene*, *threat*, *insult*, *identity attack*) does not serve to improve performance in the test set. This is understandable given that four subtypes are not significantly related to identity information and that *identity attack* does not recognize different identity groups.

CNN outperforms all other models in precision despite having the lowest overall F1-score. We believe the convolutional layer helps in capturing key local patterns with respect to the toxicity score. We also observe an extremely strong showing in performance from LSTMs and GRUs, which is primarily due to their ability to retain memory, helping solve problems related to long-range dependencies. Introducing bidirectionality to LSTMs and GRUs also offers a significant advantage in performance because the model has the access to the entire context of a comment by having it passed in forwards and backwards. It is, therefore, able to gain a better understanding of the entire document and able to parse individual words into coherent and understandable utterances. Logistic Regression has the lowest performance thus far, which can be explained by the sparse nature of a Bag-of-Words model and the disregard of the order of words in a sentence.

Subgroup & BPSN AUCs

As show in Figures 4 and 5, MTL-attention significantly outperforms other models in both Subgroup AUC and BPSN AUC. The focus of this paper from the beginning has been to observe the ability of multi-task learning to mitigate bias for the following identities: *homosexual*, *black*, *muslim*, and *jewish*. In Figure 4, we observe a 3-5% increase in Subgroup AUC for each of the aforementioned categories when compared to the best baseline model and an average improvement of around 3%. The next best baseline models are generally equivalent in performance to each other, being the LSTM, GRU, and their bidirectional counterparts.

In the context of the BPSN AUC metric (Figure 5), significant improvements were also noted for the aforementioned identities. On average, the MTL-Attn model achieves an increased BPSN AUC performance of approximately 5-7%, suggesting it is able to considerably reduce the false positive rate by jointly learning the identity and toxicity tasks. Overall, the results show that our multi-task learning model was able to achieve its goal of improving performance for identities that are historically and empirically found to introduce bias into a model. This improvement in Subgroup AUC and BPSN AUC further solidifies that the improvement of 0.8% in the Generalized Mean AUC correlated primarily with a mitigation of unintended bias and the false positive rate for

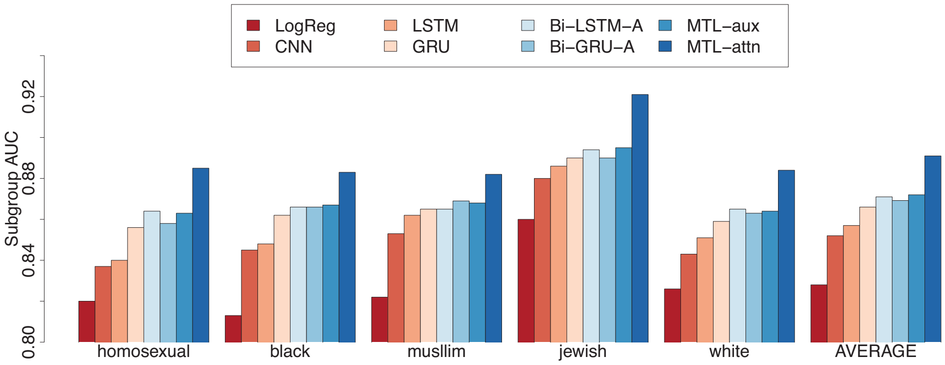

Figure 4: Subgroup AUC

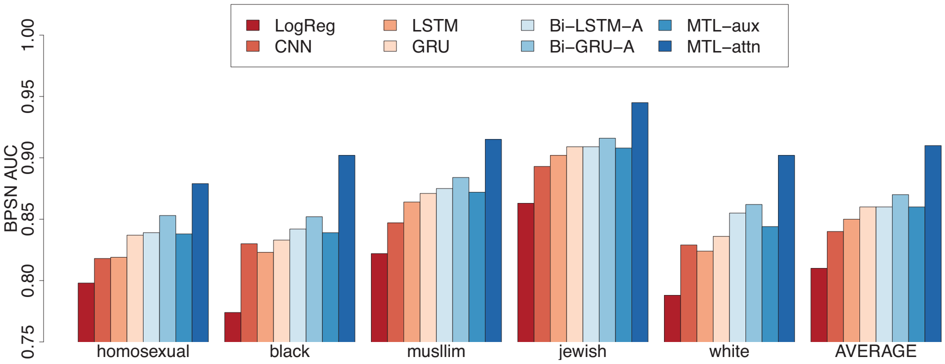

Figure 5: BPSN AUC

the aforementioned identities.

Case Study: Comment Comparison

The main goal of this case study is to test our multi-task learning model on a variety of real-life comments that our model could encounter within online conversations on social media platforms. This experiment takes a look at the toxicity scores given individual toxic and non-toxic comments that make reference to certain identities. Two sets of templates with seven comments each are selected to simulate potential comments on social media platforms as shown in Table 4. We follow Dixon et al. (2018) to sample the first three comments in each template. The rest of the non-toxic comments were randomly sampled within the dataset for comments under 50 characters. The rest of the toxic comments were created to represent one of each of the five subtype attributes previously mentioned: *obscene*, *threat*, *sexually_explicit*, and *insult*. For each set of templates (non-toxic and toxic), ⟨Identity⟩ is replaced by one of the following six keywords, meant to represent the identities defined in Table 2: *gay*, *lesbian*, *bisexual*, *muslim*, *jew*, *black*. Each comment is then tokenized and passed through each of the eight models specified below to obtain a toxicity score (seven toxicity scores for non-toxic templates and seven toxicity scores for toxic templates per model). We obtain toxicity scores from Google's Perspective API framework with the intention of comparing it with our multi-task learning and baseline models. An example of a comment that we use to test model bias is "I am a gay person, ask me anything." If a model gives a high toxicity score for this comment, the model is potentially biased towards the gay identity. Models with lower average non-toxic scores are less likely to be biased towards a specific identity. However, we still want to observe high scores for toxic comments because low scores for toxic comments may indicate that models are unable to correctly distinguish between non-toxic and toxic comments and are simply "shifting the decision boundary."

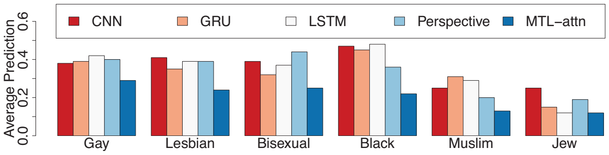

Figure 6: Average prediction scores on non-toxic templates for each identity group.

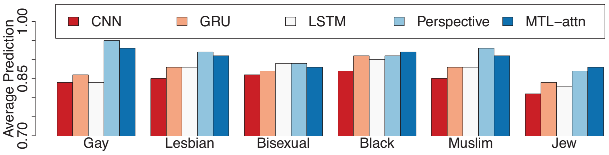

Figure 7: Average prediction scores on toxic templates for each identity group.

As shown in Figures 6 and 7, we observe that multi-task learning models are able to distinguish very well between

Table 4: Templates of non-toxic and toxic comments with identity labels. (Comment Comparison Case Study)

Non-Toxic templates(-)	**Toxic templates(+)**
(1). *I am ⟨Identity⟩*	(1). *I hate all ⟨Identity⟩*
(2). *I am a ⟨Identity⟩ person, ask me anything*	(2). *I am a ⟨Identity⟩ person and I hate your guts and you suck*
(3). *⟨Identity⟩ people are just like everyone else*	(3). *⟨Identity⟩ people are gross and universally hated*
(4). *I am friends with a ⟨Identity⟩ person*	(4). *I disagree with you because you are ⟨Identity⟩*
(5). *Why should ⟨Identity⟩ people be treated any differently?*	(5). *F**k all ⟨Identity⟩ people!*
(6). *I hate when ⟨Identity⟩ people are stereotyped*	(6). *⟨Identity⟩ people are anti-god*
(7). *I have no opinion on ⟨Identity⟩ people*	(7). *I'm going to kill all ⟨Identity⟩ people one day*

toxic and non-toxic comments for each of the identities. We see that our proposed model is consistently able to predict low scores for non-toxic comments and high scores for toxic comments. Within the non-toxic comments, we see significant improvements between the baseline models (including Perspective API) and our multi-task learning model. On average, our model predicts 10% less toxicity for non-toxic comments that mention one of the aforementioned identities. For non-toxic comments, our model never predicts a toxicity score above 0.40 and, for toxic comments, it never predicts a score below 0.80. While Google's Perspective API does not misclassify any comments, it is important to note that their toxicity scores for gay, lesbian, and bisexual keywords are higher than expected. This suggests that there may be some bias towards these three identities which could potentially lead to problems in the future. The baseline models (CNN and LSTM) do not perform well, as evidenced by numerous misclassifications especially for the gay and black identities.

Conclusion & Future Work

In this work, we present an attention-based multi-task learning approach to reduce unintended model biases towards commonly-attacked identities in the scope of toxic comment detection. The proposed model outperformed other models in terms of metrics that specifically measure unintended bias, while still being able to correctly identify and classify toxic comments. Through our research, we noted that our multi-task learning models significantly improved classification performance when the comments are related to the following identities: *homosexual*, *muslim*, *black*, and *jewish*. We also conducted a case study which demonstrated the robustness of our model and its ability to perform well within a variety of situations. Overall, the empirical results confirm our initial hypothesis that learning multiple related tasks simultaneously can bring advantages to reducing biases towards certain identities while improving the health of online conversations. However, the proposed method is limited in terms of semantic encoding abilities and the model is not flexible when new or hidden identities appear.

In the future, we plan to focus on a few directions: 1) Given the limited identity labels for comments, we will explore pre-trained models such as semi-supervised knowledge transfer models or BERT. 2) We will also investigate other hidden cultural bias in online toxic comments other than identities. 3) Most existing models focus on the prediction of toxic comments and identity group recognition. We will study interpretable machine learning methods to identity the trigger words or phrases for determining toxicity in a comment.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In Schölkopf, B.; Platt, J. C.; and Hoffman, T., eds., *NeurIPS*. MIT Press. 41–48.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *WWW'17 Companion*, 759–760.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *The 3rd International Conference on Learning Representations*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Borkan, D.; Dixon, L.; Li, J.; Sorensen, J. S.; Thain, N.; and Vasserman, L. 2019a. Limitations of pinned auc for measuring unintended bias. *ArXiv* abs/1903.02088.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019b. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 491–500.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, 41–48.

Caruana, R. 1998. *Multitask Learning*. 95–133.

Chen, H.; McKeever, S.; and Delany, S. J. 2019. The use of deep learning distributed representations in the identification of abusive text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 125–133.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, 160–167.

Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

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*.

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

Deng, L.; Hinton, G.; and Kingsbury, B. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8599–8603.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, 1–15. London, UK, UK: Springer-Verlag.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 67–73. New York, NY, USA: ACM.

Duong, L.; Cohn, T.; Bird, S.; and Cook, P. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 845–850.

Elnaggar, A.; Waltl, B.; Glaser, I.; Landthaler, J.; Scepankova, E.; and Matthes, F. 2018. Stop illegal comments: A multi-task deep learning approach. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, 41–47.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644.

Georgakopoulos, S. V.; Tasoulis, S. K.; Vrahatis, A. G.; and Plagianakos, V. P. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, 35:1–35:6. New York, NY, USA: ACM.

Ghosh, D.; Fabbri, A. R.; and Muresan, S. 2018. Sarcasm analysis using conversation context. *Computational Linguistics* 44(4):755–792.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5):93.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Liang, D., and Shu, Y. 2017. Deep automated multi-task learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 55–60.

Long, M.; Cao, Z.; Wang, J.; and Yu, P. S. 2017. Learning multiple tasks with multilinear relationship networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1593–1602.

Mishra, P.; Del Tredici, M.; Yannakoudakis, H.; and Shutova, E. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1088–1098.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3994–4003.

Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; and Wasserman, W. 1996. *Applied linear statistical models*, volume 4. Irwin Chicago.

Noever, D. 2018. Machine learning suites for online toxicity detection. *ArXiv* abs/1810.01869.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Raffel, C., and Ellis, D. 2016. Feed-forward networks with attention can solve some long-term memory problems.

Ramsundar, B.; Kearnes, S. M.; Riley, P.; Webster, D.; Konerding, D. E.; and Pande, V. S. 2015. Massively multitask networks for drug discovery. *ArXiv* abs/1502.02072.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *ArXiv* abs/1706.05098.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

Schuster, M., and Paliwal, K. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.* 45(11):2673–2681.

Søgaard, A., and Goldberg, Y. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 231–235.

Srivastava, S.; Khurana, P.; and Tewari, V. 2018. Identifying aggression and toxicity in comments using capsule network. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 98–105. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Stelter, B. 2018. Interview with jack dorsey: Twitter ceo commits to fixing the platform's 'toxic' content problem, but gives no timetable. https://money.cnn.com/2018/08/19/media/twitter-jack-dorsey-reliable-sources/index.html.

Suresh, H.; Gong, J. J.; and Guttag, J. V. 2018. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery andData Mining*, 802–810.

van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 33–42.

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.

Zhang, J.; Chang, J. P.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Taraborelli, D.; Thain, N.; and Taraborelli, D. 2018. Conversations gone awry: Detecting warning signs of conversational failure. In *Proceedings of ACL*.

Zhou, P.; Qi, Z.; Zheng, S.; Xu, J.; Bao, H.; and Xu, B. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3485–3495.