Say 'YES' to Positivity: Detecting Toxic Language in Workplace Communications

Meghana Moorthy Bhat [†] **Saghar Hosseini**[‡] **Ahmed Hassan Awadallah**[‡]
Paul N. Bennett[‡] **Weisheng Li**[§]

[†]The Ohio State University [‡]Microsoft Research [§]Microsoft
bhat.89@osu.edu, {sahoss, hassanam, pauben, weishli}@microsoft.com

Abstract

Warning: *this paper contains content that may be offensive or upsetting.*

Workplace communication (e.g. email, chat, etc.) is a central part of enterprise productivity. Healthy conversations are crucial for creating an inclusive environment and maintaining harmony in an organization. Toxic communications at workplace can negatively impact overall job satisfaction and are often subtle, hidden or demonstrate human biases. The linguistic subtlety of mild yet hurtful conversations has made it difficult for researchers to quantify and extract toxic conversations automatically. While offensive language or hate speech has been extensively studied in social communities, there has been little work studying toxic workplace communications. Specifically, the lack of corpus, sparsity of toxicity in enterprise emails and a well-defined criteria for annotating toxic conversations have prevented researchers from addressing the problem at scale. We take the first step towards studying toxicity in workplace communications by providing (1) a general and computationally viable taxonomy to study toxic language at workplace (2) a dataset to study toxic language at workplace based on the taxonomy and (3) analysis on why offensive language and hate-speech datasets are not suitable to detect workplace toxicity. Our implementation, analysis and data will be available at https://aka.ms/ToxiScope.

1 Introduction

Studies have shown that more than 80% of the issues affecting employees' productivity and satisfaction are related to negative work environment behaviors such as harassment, bullying, ostracism, gossiping, and incivility (Anjum et al., 2018). Moreover, workplace gossiping results in distracted employees and low morale.Duffy et al. (2002) and Kong (2018) find that workplace incivility leads to social undermining of employees which could lead to trust issues, difficulty in establishing cooperative relationship, lower job satisfaction and attitudinal outcomes such as gaining personal power and reputation (Aquino and Thau, 2009; Baumeister, 1995; Ellwardt et al., 2012; McAndrew et al., 2007).

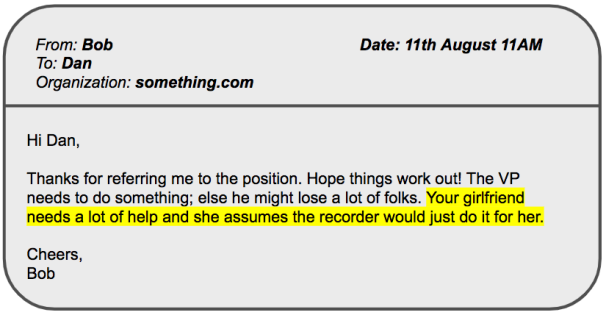

Figure 1: An example of workplace communication. The highlighted sentence was annotated as toxic and gossip by annotators. This instance has a confidence score of 0.15 on Perspective API[1]

Many organizations enact policies that prohibits practicing extremely toxic behaviors like bullying, verbal threats, profanity, harassment and discrimination; yet detecting more subtle forms of toxicity like negative gossiping, stereotyping, sarcasm, and microaggressions in conversations remains a challenge.

Toxicity can be manifested in different ways. It spans a wide spectrum that includes subtle and indirect signals; that can often be no less toxic than overly offensive language (Jurgens et al., 2019). While the research community has made enormous progress in detecting overly offensive language and hate speech (Schmidt and Wiegand, 2017; Waseem et al., 2018; Fortuna and Nunes, 2018; Qian et al., 2019), there has been less focus on computationally evaluating other subtle expressions of toxicity.

Most of the work was done while the first author was an intern at Microsoft Research

[1] https://www.perspectiveapi.com

Qualitative studies have found these subtle signals to have long lasting negative effect (Sue, 2010; Nadal et al., 2014). As Figure 1 shows, currently popular toxicity detection tools cannot detect subtle yet hurtful conversations as harmful. We argue that it is equally important to detect these subtle aggressive conversations and educate employees for a healthy workplace. Detecting wider aspects of toxic text can be challenging. Subtle signals like stereotyping, mild aggression can be context-sensitive, sparse, highly subjective and do not have well defined annotation guidelines; whereas overly toxic language and hate speech are rarely context-sensitive (Pavlopoulos et al., 2020) and have well-defined guidelines (Waseem et al., 2017). In this paper, we take first steps towards (1) defining a taxonomy for studying toxic language in workplace setting by analyzing the definitions from impoliteness theory and psychology (2) building a dataset of human annotations on publicly available email corpus (3) providing computational methods to establish baselines for detecting toxic language in enterprise emails, and (4) analyzing why current datasets and tools for detecting hate speech do not work in our setting.

2 Related Work

Offensive Language Detection: Perspective API is a popular toxicity detector for detecting offensive conversations. Waseem et al. (2018) devised a taxonomy and created a dataset to detect hate speech and discrimination. Xu et al. (2012) studied bullying, Chatzakou et al. (2017) released a dataset to study bullying in online posts, and Zampieri et al. (2019a) released a corpus for offensive posts named OffensEval which has been encouraging researchers to study offensive contents. Recently, Safi Samghabadi et al. (2020) released a dataset with emojis for identifying sexually profane language and Rajamanickam et al. (2020) showed joint model of emotion and abusive language detection helps model performance. However, toxic language in workplace has often subtle aggressive conversations and lesser offensive text. Subtle aggressive conversations can be covert faux pas or unintentional whereas offensive text is overt and includes intentional choice of words. Also, a conversation in a workplace is more formal than the social media text. Due to their fundamental different structure, current datasets and models trained on these datasets are not able to properly detect workplace toxicity.

Microaggression datasets: Breitfeller et al. (2019) released a dataset from Reddit, Gab, and www.microaggressions.com showing that it's possible to annotate these highly subjective and linguistically subtle uncivil communications and detect them using computational methods. It is focused on gender-based discrimination due to their availability in social media. The annotation guideline also use gender as discrimination axis to determine toxicity. Whereas we are interested in formal conversations that are context dependent and are majorly targeted towards individuals addressed in emails irrespective of gender. Wang and Potts (2019) introduced a new Reddit dataset with labels corresponding to the condescending linguistic acts in conversations and showed that by leveraging the context, it is possible to detect this type of challenging toxic language. Similarly, Caselli et al. (2020) leveraged the context of occurrences to create a Twitter dataset for implicit and explicit abusive language. Implicit abusive language does not immediately insinuate abuse. However, its true meaning is often concealed by lack of profanity or hateful terms which makes it difficult to detect. Oprea and Magdy (2020) released a corpus for sarcasm self-annotated by authors on Reddit. However, these datasets mainly contain abusive language and sarcastic tweets on popular social events and are informal.

To the best of our knowledge, there is no available dataset in our community to study toxic language in emails. The most similar work to ours can be Raman et al. (2020). However, the focus of this work has been mostly offensive language in GitHub community whereas our work focuses on detecting toxicity in workplace emails.

Email Communications: There is also some prior work on Email corpus for sociolinguistic downstream tasks. Prabhakaran et al. (2014) explored the relation between power and gender on Enron corpus. They showed that the manifestations of power differ significantly between genders and the gender information can be used to predict the power of people in conversations. Similarly, Bramsen et al. (2011) studied social power relationships between members of a social network, based purely on the content of their interpersonal communication using statistical methods. Madaan et al. (2020) released automatically labeled Enron corpus for politeness. However, their definition for polite-

ness does not capture toxic language. Chhaya et al. (2018) devised computational method to identify conversation tone in Enron corpus. They categorize tones as frustration, formal and polite and find that affect-based features are important to detect tone in conversation. However, affect-based features do not capture subtle offensive text. We are interested in studying subtle and offensive text in workplace emails which are different from the prior work in this area.

3 Toxicity in Enterprise Email

Our goal is to study and understand workplace toxic communications through one of the most frequently used ways of communication in organizational settings, emails (The Radicati Group, 2020). The distribution of our dataset (Section 3.2) demonstrates the significant presence of the implicit and subtle toxic language in workplace email communications contrary to social media and open source communities. Table 1 also provides the statistics of different datasets that study the implicit and explicit toxic language.

Dataset	size	toxic comments	Type	Agreement Score
(Raman et al., 2020)	1594	189 (11%)	Explicit	N/A
(Breitfeller et al., 2019)	1065	337 (30%)	Implicit	0.41
(Wulczyn et al., 2017)	69.5k	26.5k (37.4%)	Explicit	0.45
ToxiScope (Ours)	10k	1210 (11.9%)	Implicit	0.77

Table 1: Distribution of different datasets that study implicit and explicit toxic language.

We created a taxonomy (Section 3.1) and a crowd sourced annotation task (Section 3.2) to manually annotate toxic language in the Avocado research email collection (Oard et al., 2015). This collection contains corporate emails from an information technology company referred to as "Avocado". The collection contains an anonymized version of the full content of emails, and different meta information from Outlook mailboxes of employees' emails. The full collection contains 279 employees and 938,035 emails.

In addition, we perform analysis of different emotional affects for each category of toxic language. From previous work, we understand that toxic language has a strong correlation with negative emotions. We also studied whether using context was beneficial in determining toxicity. To this end, we conducted an analysis to study whether humans benefit from context in detecting toxic language in emails. We assume that to determine toxicity in a text, humans read the entire email body and previous emails and not only the given text. We quantify these observations through annotations before using context aware representations in our modeling.

3.1 Taxonomy for toxic language

We leveraged the different negative culture practices with definitions from impoliteness theory (Culpeper, 1996) and offensive language detection in social media (Zampieri et al., 2019b,a) to define taxonomy for toxic language in workplace communications. We have the following goals in mind: (1) generalizable across different organizations, (2) sufficiently represented in our corpus, (3) cover the main dimensions of negative culture in workplace from cross-domain literature. We have summarized the definitions in Table 2 and described each of these below.

Non-Toxic: The non-toxic class has instances of friendly, knowledge sharing, formal respectful type of conversations. These conversations often have positive or neutral connotations.

Impolite: The impolite class has instances of sarcasm, stereotyping, rude statements. These conversations often have opposite polarity to their previous context with negative or neutral connotations that might complement the work on benevolent sexism (Jha and Mamidi, 2017). Following Impoliteness theory (Culpeper, 1996), we define 'Rude' as direct, intentionally disrespectful words to the addressee whereas sarcasm (implicature to express the opposite of being said), stereotyping (unintentional) need not be necessarily direct yet disrespectful comments to the addressee in the conversation.

Negative Gossip: The gossip class includes rude, mocking conversations about a person not involved in the conversation. We find these instances have negative connotations with a tone of complaint and lack of respect toward the target. Kong (2018) found repeated gossip conversations in organizations caused hostility and stress among the employees. As shown by Wulczyn et al. (2017), conversations targeted targeted towards a third person need not necessarily be extreme yet can be disrespectful. Evidently, our annotators find our annotators feel gossip conversations are more annoying whereas impolite conversations have more sadness with higher overlap with offensive category (Figure 3). We refer to this type as "Gossip" in the rest of the paper.

Offensive: Detecting overly toxic language has

Type	Sub-type	Example
Non-toxic	NA	Hey, how are you holding up?
		Can you please reschedule the meeting for tomorrow?
Impolite	sarcasm	You need big glasses huh, LOL!!!? Its 11:00AM
	stereotype	Ladies, since you all are good at cooking and are used to it,
		I invite you to participate for potluck in office.
	forced teaming	We all are victims of the new policy. Let the retaliation begin!
	authoritarian	I want you to give me the numbers by 9PM today. I do not have time to wait until tomorrow.
	rude	I did not want to yell at you in front of everyone, but you are performing poorly!
Negative Gossip	mocking	When I take a long time I am slow and when my boss takes a long time, he is thorough
	complain	How does this guy function in society?
Offensive	profanity	Let's kiss their a** and get it done.
	discrimination	Would you rather be called African-American or black?
	bullying	Whoever is doing these tags is brain dead enough to send the wrong tag.
	violence	All [nationality/race] are lazy and don't deserve to work here
	harassment	Your backside is banging in that dress.

Table 2: Sub-type categories of toxic language that we developed based on the literature, and email conversations. Examples demonstrate that the phenomenon is complex and is different from offensive text or negative sentiment.

been extensively studied in the research community. We follow a similar definition of offensive language as Zampieri et al. (2019b) which refers to any form of unacceptable language to insult a targeted individual or group. In our setting, we define offensive language such that it includes five broad categories: profanity, bullying, harassment, discrimination and violence.

3.2 Annotation task

We design a hierarchical annotation framework to collect instances of sentence in an email and the corresponding label on a crowd-sourcing platform. Before working on the task, annotators go through a brief set of guidelines explaining the task. We collect the dataset in batches of around 1000 examples each. For the first three batches, we upload 75-100 instances manually labeled as toxic by the group of researchers working on the project to understand if the annotators followed the guidelines. We repeat the pilot testing until desirable performance is achieved. Also, we manually review a sample of the examples submitted by each annotator after each batch and exclude those who do not provide accurate inputs from the annotators pool and redo all their annotations. A key characteristic of subtle toxic emails are that they often result from prior experiences, cultural difference or background between individuals (Sue et al., 2007). Hence, designing annotation for detecting toxicity is a difficult task and there will be discrepancies in perceived toxicity between the annotators. In order to minimize ambiguity and provide a clearer context to the annotators, we provide email body, subject, and the prior email in thread as context information. For each highlighted sentence, annotators indicate whether the post is toxic, type of toxicity, whether the target of the toxic comment is the recipient or someone else, whether the prior email as context was helpful, the kind of negative affect associated with toxicity and whether the whole email was toxic. We provide a subset of negative affects to the annotators from WordNet-Affect (Strapparava and Valitutti, 2004). The annotators answer the questions on type of toxicity and the target only if they indicate potential toxicity during annotation. They can also choose multiple toxic categories for a highlighted sentence. Finally, the annotators are provided an optional text box to provide additional details if the highlighted sentence did not belong to any of the categories we defined. Please note that the sub-types of toxicity do not have a clear boundary and are not mutually-exclusive.

A total of 76 annotators participated in this task. All annotators were fluent in English and came from 4 countries: USA, Canada, Great Britain and India, with the majority of them residing in the USA. Each highlighted statement in the email was annotated by three annotators and they were compensated based on an hourly rate (as opposed to per annotation) to encourage them to optimize for quality. They took an average of 5 minutes per annotation. We assume a sentence is toxic even if one out of three annotators perceived it as toxic. We adopt this principle to be inclusive of every individual's background, culture, sexual orientation and implicit toxic language can be subtle. Similarly, we included the union of the toxicity types selected by the three annotators for the instance. A snapshot of our crowd-sourcing framework can be found in

Appendix 5

Due to the scarce nature of toxic conversations in emails, we adopt two round approach for data collection. For the first round of annotations, we use several heuristics to increase the chances of identifying positive instances in the sample. We tried running the Perspective API and the microaggression model (Breitfeller et al., 2019) against Avocado corpus. The coverage of Perspective API is extremely low (0.1%) since not many overly toxic text is present in Avocado corpus. On the other hand, the microaggression model output has low precision (0.12%). To further prune the false positives, we employ filtering methods[2] over the outputs from microaggression model before sending the positive labels for annotation. The first round of annotations provided a positive label ratio of 2.74% compared to 0.29% from a manually annotated batch of around 800 random email sentences. This implies the need to be selective regarding the emails we submit for annotation. In addition, for the second round of annotations, we used SVM classifier to pick positive instances from the unlabeled email corpus. To avoid model biases, we randomly sample unlabeled email sentences based on their probability scores with more instances being sampled from the higher scores ranges. The second round of annotations provided a positive label ratio of 11.2% which is significantly higher than our previous rounds.The classifier is updated with more examples after each round of annotations.

Overall, the final dataset contains 10,110 email sentences of which 1,120 of the sentences are labeled as toxic by annotators. We call this dataset for studying toxic language in workplace communications as ToxiScope. Please note that we asked the annotators to identify spam emails and their types including Advertisement, Adult content, and Derogatory content. We observed that 99% of the emails in Spam category are advertisement and we decided to exclude those emails since advertisement contents are not in the scope of toxic language detection. Figure 2 shows the distribution of toxic emails over sub-categories of toxic language which indicates higher frequency for Impolite emails.

Annotators Agreement: Overall, the annotations showed inter-annotator agreement score of Krippendorf's $\alpha = 0.718$ to detect whether a given sentence was toxic or not. Broken down by each category, annotators agreed on a sentence being offensive at Krippendorf's $\alpha = 0.77$, impolite at Krippendorf's $\alpha = 0.29$ and gossip at Krippendorf's $\alpha = 0.32$. The high agreement score on overall toxicity shows that annotator judgements are reliable and the lower agreement score on sub-types are indicative of the subjectivity and lack of objectivity for implicit toxicity (Lilienfeld, 2017) and not the quality. We also quote several prior works in toxicity setting and other tasks that lack objectivity, and have inter-annotator agreement score in our range. Microaggression dataset has a score of 0.41 for 200 instances and Rashkin et al. (2016) has a score of 0.25 for inter-annotator agreement.

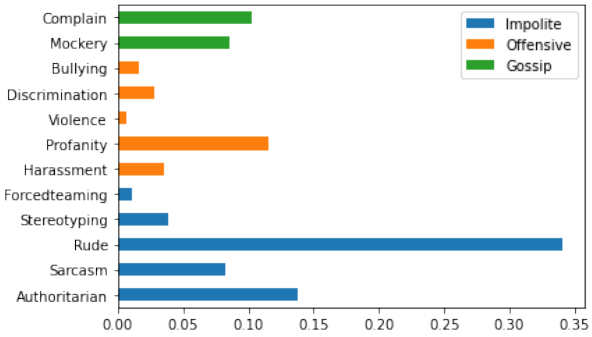

Figure 2: Frequency of each sub-category of toxic sentences.

Insights from annotation task: Sometimes defining a clear boundary between categories of toxic language is challenging because they are not mutually exclusive. Therefore a statement can belong to multiple toxic categories. For example, the content of an email can be about gossiping and at the same time be discriminatory against a certain group of people. Our analysis shows that 92% of emails belong to a single toxic category while the rest of the emails contain two or more types of toxic language. Figure 3 shows the co-occurrence of different toxic contents in the same email. We can observe that the Offensive and Impolite categories are slightly more likely to happen in the same email than with Gossip. Since our task is highly subjective, in order to understand the reasons behind perceived toxicity we ask annotators several questions about the target and affect of the toxic statement, and whether the context (previous email) is useful in determining the toxicity of the statement. We find that in 41% of the instances, context information was helpful to determine toxicity. In 76.86% of the toxic instances, the language was targeted to another individual or a group. Un-

[2]LIWC lexicon (Pennebaker et al., 2015), WordNetAffect (Strapparava and Valitutti, 2004), https://github.com/snguyenthanh/better_profanity

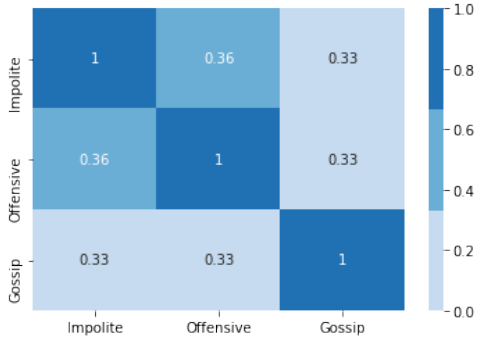

Figure 3: Correlation between emails toxic categories.

derstandably, all the toxic instances have negative affect with anger and hostile being present in most of the cases. However, annotators find gossip examples more disgusting and a toxic sentence to be 6.1% more annoying when they are targeted to another individual not in the conversation.

We use 70% of the data for training and 10% as validation set. We hold out 20% of the data for test set. Table 3 provides a summary of the final dataset.

Sentence Type	Train	Dev	Test
Toxic	886	117	207
Impolite	636	84	139
Gossip	176	23	47
Offensive	74	10	21
Non-toxic	6308	864	1728
Total	7194	981	1935

Table 3: Number of instances in each toxic category and set of ToxiScope

4 Detecting toxic conversations in Emails

We design our experiments with the following goals: (1) Investigate if contextual information (email body, the parent email) helps in determining toxicity. We also study which categories of toxic language benefit from adding context to the sentence. (2) We also test our hypothesis that current toxic language datasets cannot identify indirect aggressive or impolite sentences. We consider current state-of-the-art toxic language detectors for this task. (3) Evaluate our baseline models on other datasets including Wiki Comments (Wulczyn et al., 2017) and GitHub (Raman et al., 2020) to study if understanding subtle signals help in determining overly toxic language.

We experimented with publicly available state-of-the-art models in literature and the Perspective API:

Linear Models: We generate n-grams (where n is up to 2) and feed them as feature vectors for the classifier. We experiment with Logistic Regression and Support Vector Machines (SVM) as utilized by Breitfeller et al. and Raman et al. for our task.

Context-Aware Sentence Classification: Wang et al. developed a GRU model with context encoder that uses attention mechanism on the context sentences and a fusion layer that concatenates target and context sentence representations to study the influence of context in intent classification. We leverage this model for our experiments.

Bert Classification: We experimented with the Bert-based model proposed by Liu et al.. We fine-tuned the model that was initially trained on Zampieri et al. with ToxiScope. This model concatenates the text of the parent and target comments, separated by Bert's [SEP] token, as in Bert's next sentence prediction pre-training task.

Bert+ MLP: For this model, we experimented with context-aware version of Bert-based classifier as explained above. We freeze the first 8 layers of Bert and add a non-linear activation function before the classification layer.

5 Results and Analysis

Table 4 summarizes the performance of models trained and tested on ToxiScope. The baselines performance are reported for binary classification (toxic vs non-toxic). We report evaluation metrics in F_1(macro and micro) and accuracy (TPR and TNR) of different classes due to class imbalance. For the models in Table 4, which required context as an input, we took the prior email in the thread during pre-processing. The results imply pretrained Bert models fine-tuned on ToxiScope perform better than non-pretrained models. Hence, we will focus on these models to evaluate the effect of context on the outcome. In addition, the low recall performance or True Positive Rate (TPR) demonstrates the challenge in detecting subtle toxic instances in communications and from now on we pay more attention to TPR and F_1 score metrics.

Effect of adding context: As outlined in Section 3.2, annotators find prior email and email body helpful to determine toxicity. Pavlopoulos et al. (2020) showed that adding context did not help pre-trained models like Bert in boosting the

Model	Accuracy			F_1
	toxic (TPR)	*non-toxic (TNR)*	*overall*	*(macro/micro)*
Logistic Regression	0.0097	1.00	0.5050	0.4816/0.8941
Linear SVM	0.3092	0.9421	0.6257	0.6378/0.8744
DCRNN (Wang et al., 2019)	0.1223	1.00	0.5610	0.4980/0.8537
Bert Classification	**0.4348**	0.9825	0.7102	0.75/0.91
Bert + MLP	0.4300	0.9925	**0.7112**	**0.7696/0.9213**

Table 4: Performance of different models trained and tested on ToxiScope. We report True Positive Rate (TPR), True Negative Rate (TNR), and overall accuracy along with F_1 (macro and micro) scores.

Model	Context	Offensive	Gossip	Impolite	Average
Bert Classification	no context	0.75	0.3333	0.3089	0.4640
	email body	0.675	0.2410	0.3581	0.4247
	(+/-1) adjacent sentences	**0.80**	0.5027	0.2133	0.5053
	previous email	**0.80**	0.3675	**0.3966**	0.5213
Bert + MLP	no context	0.75	0.39	0.379	0.5063
	email body	0.75	0.4718	0.375	0.5322
	(+/-1) adjacent sentences	**0.80**	**0.5156**	0.1869	0.5008
	previous email	**0.80**	0.4523	0.365	**0.5391**

Table 5: Performance of our baseline models across different categories of toxic language. We report True Positive Rate (TPR) for each category and the average over their TPR.

performance. However, the dataset in their setting was small in size and the target comments were mostly offensive. These observations may not generalize in our case since we are interested in detecting implicit and subtle cases of aggressive language. In order to evaluate the effect of the contextual information, we experimented with different variations of the context. Table 5 presents the TPR for different categories of the toxic language. Based on our experiments, models find context helpful to detect toxicity. Interestingly, models do not find contextual information necessary to detect offensive language unlike other categories. We also observed gossip category benefits the most from the neighborhood sentences as context. The majority of the gossip emails in our dataset belong to complain sub-category which are spread across multiple sentences. Hence, many of the neighboring sentences could have had negative connotations that would have aided the models. However, on average using the previous email in the thread is most helpful in detecting the toxic language. In general, finding implicit toxic language is a difficult task. This is evident in low TPR of gossip and impolite classes as well as their sparse labels and the low inter-annotator agreement scores in those categories.

Generalization to other domains: To investigate how other domains can lever our dataset, we trained the baseline models for toxic language detection (Breitfeller et al., 2019; Raman et al., 2020) and context aware sentence classification (Wang et al., 2019) on ToxiScope. Then, we tested these models against different toxic language datasets. Since we did not find any dataset studying toxic language in workplace (with implicit and explicit toxic text), we picked the datasets that overlap with one or few categories of our interest. The results are presented in Table 6 which shows that Bert based models outperform other methods in all of the domains. Note that on microaggression dataset we achieve TPR of 0.54 which performs better than the model provided by Breitfeller et al. (2019) with best TPR of 0.36[3]. On Wiki Comments dataset, our baseline models using Bert have good accuracy (TPR 0.86) in detecting toxic text which is comparable to the TPR of Perspective API (0.85). The reason for high false positive rate could be that Wiki Comments dataset does not consider subtle

[3]Since test set for Microaggression datset is not publicly available, we randomly split the available set to 80:20 for training and test.

Model	Microaggression dataset		Wiki Comments		GitHub	
	F_1	*Accuracy*	F_1	*Accuracy*	F_1	*Accuracy*
	(macro/micro)	*toxic (TPR)*	*(macro/micro)*	*toxic (TPR)*	*(macro/micro)*	*toxic (TPR)*
Logistic Regression	0.4169/0.6769	0.014	0.6451/0.7964	0.4903	0.3413/0.5181	0.0
Linear SVM	0.5427/0.6056	0.3571	0.4867/0.5668	0.5870	0.4751/0.5544	0.1720
DCRNN (Wang et al., 2019)	0.4517/0.6914	0.13	0.5215/**0.8856**	0.2382	0.3997/0.5231	0.051
Bert Classification	**0.6578/0.7136**	0.4714	**0.7430**/0.8388	0.7805	0.4368/0.5506	0.1011
Bert+MLP	0.6233/0.6573	**0.5429**	0.7210/0.8070	**0.8608**	**0.5525/0.5843**	**0.2287**

Table 6: Performance of baseline models trained on ToxiScope and tested on several toxic language datasets.

aggressive text as toxic. The best performing classifier by Raman et al. (2020) on GitHub datatset has a TPR of 0.35. One reason for poor scores on GitHub dataset can be attributed to noisy labels. We sampled a few instances from GitHub dataset and found 15% of them to be noisy. Overall, these experiment results imply the potential benefits of using our dataset for detecting toxic language in social media and open source community domains.

Leveraging social media and open source communities data to detect workplace toxicity: Offensive language is widely studied on social media language and there are several datasets and methods available for this task. Tables 8 presents the performance of the publicly available models and API[4] on ToxiScope. The model from Breitfeller et al. (2019) has a reasonable performance on ToxiScope. Their method uses lexicons for microaggressions from external sources. Leveraging these external sources as weak supervision signals might help in boosting performance of models for ToxiScope as well.

Next, we investigated if these datasets can be helpful in training models for detecting workplace toxicity. We fine-tuned and trained Bert based models over Microaggression, GitHub, and Wiki Comments and ran the inference on ToxiScope. As we expected, Table 7 shows that the models trained on Microaggression dataset are more applicable to workplace toxic language detection. However, they are still performing worse than the in-domain models (Table 4). Impolite and gossip (constituting of sarcasm, stereotyping, rude) categories are predominantly present in ToxiScope while there are not many datasets available for these tasks and the existing datasets are small in size. This could explain the inadequate performance of these models.

[4]We utilized Perspective API which is trained over 160k human labeled annotations of Wikipedia comments.

6 Conclusion

Previously, we saw a gap in available resources to detect workplace negative communications and based on our observations, Microaggression dataset was the only resource applicable to this domain which did not show promising performance. Hence, we created ToxiScope to close this gap. We presented a taxonomy and annotation guidelines to study toxic language in workplace emails. We also provided baseline methods to detect toxic language in ToxiScope. Further, we demonstrated the necessity of new dataset to detect workplace toxicity since the models trained on existing overly toxic datasets and on Microaggression dataset do not detect subtle toxic text. In addition, we observed that context help Bert based models to detect subtle toxic sentences. However, our results indicate that we need more sophisticated models and better representation of context to detect implicit toxic sentences. In future, we will explore other methods like weak supervision from other sources and

Model	Microaggression		Wiki Comments		GitHub	
	F_1	*Accuracy*	F_1	*Accuracy*	F_1	*Accuracy*
	(macro/micro)	*toxic (TPR)*	*(macro/micro)*	*toxic (TPR)*	*(macro/micro)*	*toxic (TPR)*
Bert Classification	0.6780/0.8889	0.3720	0.6078/0.8992	0.1739	0.4906/0.8941	0.0483
Bert+MLP	0.6921/0.9106	0.3188	0.5951/0.8956	0.1594	0.5971/0.9070	0.1401

Table 7: Performance of Bert models trained on Microaggression, Wiki Comments, GitHub datasets and tested on ToxiScope. The column denotes the dataset all the models were trained on.

Model	Accuracy			F_1
	toxic (TPR)	*non-toxic (TNR)*	*overall*	*(macro/micro)*
Perspective API	0.2174	0.9907	0.6040	0.6432/0.8848
Raman et al. (2020)	0.1014	0.9797	0.8858	0.5492/0.8734
Breitfeller et al. (2019)	0.3987	0.5556	0.5217	0.4375/0.5483
Liu et al. (2019)	0.4348	0.9825	0.7102	0.75/0.91

Table 8: Performance of different models with inference on ToxiScope.

self-training for better performance.

Going forward, we will also investigate other research questions pertaining to the likelihood of an individual using toxic language repeatedly, correlation of power and gender dynamics with respect to toxicity, presence of the bias (racial/gender) in ToxiScope, understanding the degree of severity of toxic text. We hope our work will encourage the researchers in the community to study and develop methods to detect workplace toxicity.

Acknowledgments

We thank anonymous reviewers for their constructive feedback and Saleema Amershi, Michael Gammon, Alexandra Olteanu, Allison Hegel, Liye Fu, Subho Mukherjee for valuable discussions during the project.

7 Ethical Considerations

7.1 Annotation

In this work, we leverage the publicly available Avocado corpus which belongs to Language Data Consortium (LDC). This email dataset has been processed and anonymized by LDC. We received approval from our organization Internal Review Board (IRB) before starting the annotation task to make sure we are in compliance with the Avocado Research Email Collection license agreements as well as the ethical guidelines. We understand that annotating potentially toxic content can have negative impact on the workers. In order to reduce these effects, we provided warnings and information about the research project in a consent form. We asked the annotators to read the consent form and only proceed if they've agreed to its terms (Figure 4). The risks and benefits of working on this annotation tasks were presented to annotators in the consent form:

Benefits: There are no direct benefits to you that might reasonably be expected as a result of being in this study. The research team expects to learn to detect micro-aggressive and toxic language in email communications from the results of this research, as well as any public benefit that may come from these Research Results being shared with the greater scientific community.

Risks: During your participation, you may experience some discomfort being exposed to profanity, toxic and discriminatory language in emails. To mitigate this risk, the research team makes it possible for you to take a break or skip tasks without adversely affecting your ratings within the crowdsourcing platform. This research may involve risks to you that are currently unforeseeable.

In addition, we did not collect any personal or demographic information other than their crowd source platform identification number. The consent form explains how we manage their information and provide details about their compensations. Resources were also provide to answer the annotators questions and concerns. Moreover, we limited the number of emails an annotator can work on in a task and paid them above minimum wage ($12-15 per hour).

7.2 Deployment

Detecting harmful language in email communication is a difficult task even for human. Recent work have shown that the toxic language detection models are also very prone to racial biases (Sap et al., 2019; Davidson et al., 2019) due to the fact that they are using biased datasets. In this work, we hired annotators from different English speaking countries to reduce the bias in our dataset. However, this is a research paper with the goal to better understand the problem of toxic language in workplace communications and encouraging other researchers to work on this problem. We believe further study needs to be done on this dataset to make sure it's not biased before deploying any computational model.

In addition, for deploying this technology, we need access to the employees' communications. To the best of our knowledge, most workplaces do not provide any guarantee of privacy for employee's communications using enterprise systems. In addition, there are several existing technologies being implemented on workplace communications for improving users' productivity such as response generation and intent detection in emails. These technologies are being used without violating user's privacy thanks to advances in the fields of unsupervised learning and privacy-preserving machine

learning.

Moreover, this technology have multiple applications and some of them can potentially be used to harm employees and their friends and family. For example, using this model to detect toxic language and report employees to HR or their manager is a high-stake application. If this system makes a false positive error, it may damage employee's reputation, forces the employee to defend themselves and diminishes their trust in the company. This technology can also be used to provide feedback to employees about their written communication style. This tool can be used for training purposes and increasing workers awareness of such a micro-aggressive language. If this system makes frequent false positive errors, employees will become annoyed and be less productive, which causes an eventual drop in the company's profits. Companies can pursue mitigation steps and allow employees to provide feedback and dispute the system's predictions.

References

A. Anjum, X. Ming, and S. F. Siddiqi, A. F.and Rasool. 2018. An empirical study analyzing job productivity in toxic workplace environments. International journal of environmental research and public health.

Karl Aquino and Stefan Thau. 2009. Workplace victimization: Aggression from the target's perspective. *Annual Review of Psychology*, 60(1):717–741. PMID: 19035831.

Leary M. R. Baumeister, R. F. 1995. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497–529.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 773–782, USA. Association for Computational Linguistics.

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

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

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 13–22, New York, NY, USA. Association for Computing Machinery.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics*, 25(3):349 – 367.

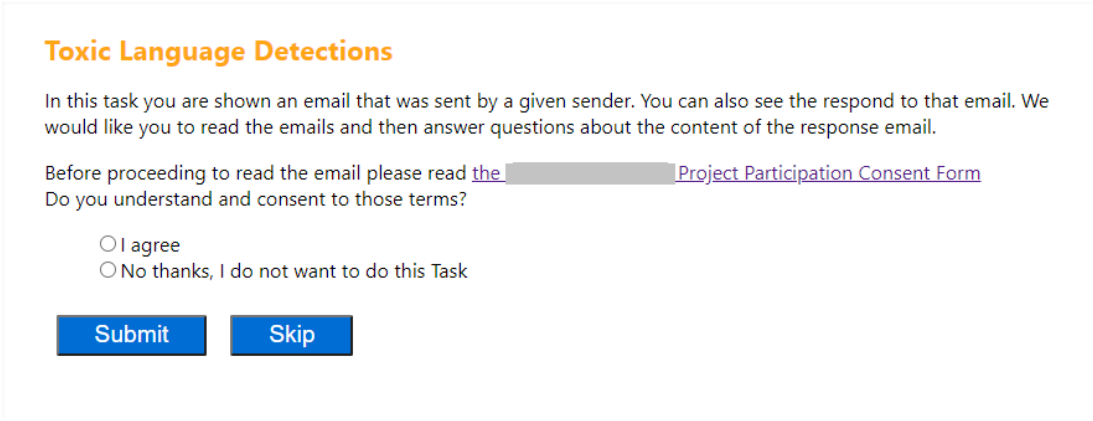

Figure 4: A snapshot of the annotation task which shows the annotator must read the consent form and agree to its terms before proceeding to annotate an email

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

Michelle K. Duffy, Daniel C. Ganster, and Milan Pagon. 2002. Social undermining in the workplace. *Academy of Management Journal*, 45(2):331–351.

Lea Ellwardt, Giuseppe (Joe) Labianca, and Rafael Wittek. 2012. Who are the objects of positive and negative gossip at work?: A social network perspective on workplace gossip. *Social Networks*, 34(2):193 – 205.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Ming Kong. 2018. Effect of perceived negative workplace gossip on employees' behaviors. *Frontiers in Psychology*, 9:1112.

Scott O. Lilienfeld. 2017. Microaggressions: Strong claims, inadequate evidence. *Perspectives on Psychological Science*, 12(1):138–169. PMID: 28073337.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach.

Francis T. McAndrew, Emily K. Bell, and Contitta Maria Garcia. 2007. Who do we tell and whom do we tell on? gossip as a strategy for status enhancement1. *Journal of Applied Social Psychology*, 37(7):1562–1577.

Kevin L. Nadal, Katie E. Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development*, 92(1):57–66.

Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado research email collection. DVD.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

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

James Pennebaker, Martha Francis, and Roger Booth. 2015. Linguistic inquiry and word count (liwc).

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Naveen Raman, Minxuan Cao, Yulia Tsvetkov, Christian Kästner, and Bogdan Vasilescu. 2020. Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions. New York, NY, USA. Association for Computing Machinery.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2020. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online. Association for Computational Linguistics.

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

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

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Derald Sue, Christina Capodilupo, Gina Torino, Jennifer Bucceri, be Aisha, Kevin Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: Implications for clinical practice. *The American psychologist*, 62:271–86.

Derald Wing Sue. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Wiley.

INC. The Radicati Group. 2020. Email statistics report, 2020-2024.

W. Wang, Saghar Hosseini, Ahmed Hassan Awadallah, P. Bennett, and Chris Quirk. 2019. Context-aware intent identification in email conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

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

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Wikipedia talk labels: Personal attacks.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, page 656–666, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

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

A Appendix

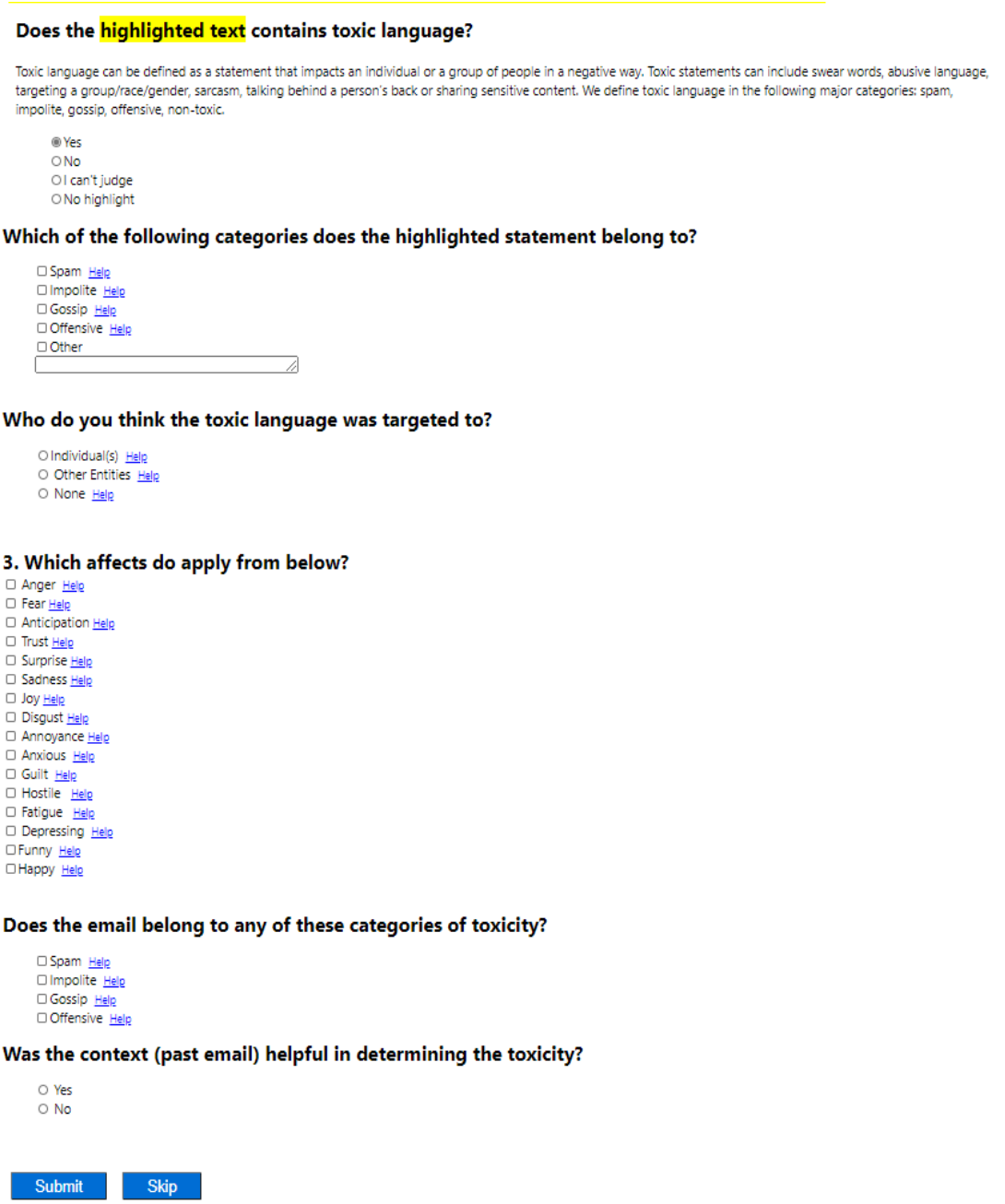

Figure 5: Snapshot of crowd sourcing task.