DeTexD: A Benchmark Dataset for Delicate Text Detection

Serhii Yavnyi* **Oleksii Sliusarenko*** **Jade Razzaghi*** **Olena Nahorna***
Yichen Mo* **Knar Hovakimyan*** **Artem Chernodub***
Grammarly
{firstname.lastname}@grammarly.com

Abstract

Over the past few years, much research has been conducted to identify and regulate toxic language.[1] However, few studies have addressed a broader range of sensitive texts that are not necessarily overtly toxic. In this paper, we introduce and define a new category of sensitive text called "delicate text." We provide the taxonomy of delicate text and present a detailed annotation scheme. We annotate DeTexD, the first benchmark dataset for delicate text detection. The significance of the difference in the definitions is highlighted by the relative performance deltas between models trained each definitions and corpora and evaluated on the other. We make publicly available the DeTexD Benchmark dataset, annotation guidelines, and baseline model for delicate text detection.[2] [3] [4]

1 Introduction

The prevalence of user-generated toxic language on online social networks has motivated many to develop automatic methods of detecting such content (Warner and Hirschberg, 2012), (Waseem and Hovy, 2016), (Davidson et al., 2017), (Schmidt and Wiegand, 2017), (ElSherief et al., 2018a), (ElSherief et al., 2018b), (Qian et al., 2018a), (Qian et al., 2018b). These efforts towards moderating toxic language have gained even more momentum as large language models, which have the potential to generate harmful content, have become more mainstream (Welbl et al., 2021), (Bender et al., 2021), (Hovy and Prabhumoye, 2021), (Kocielnik et al., 2023). Much of this work has been constrained to texts that are toxic or otherwise overtly harmful; however, there are many other sensitive texts where interaction with other users or virtual agents may be triggering or offensive. While some studies (Yenala et al., 2018), (Parnell et al., 2020), (Tripathi et al., 2019) have addressed specific sensitive areas (e.g., insults, geopolitics, or illegal activity), to our knowledge, this is the first study that comprehensively analyzes sensitive content in general.

Text	Delicate	Hate speech	Offensive	Profanity
This is f*cking amazing!	no	no	no	yes
Sometimes I have suicidal thoughts but I never talk about it with my mom.	yes	no	no	no
I think you are not a good person and I don't need your toxicity in my life.	yes	no	yes	no
You are full of sh*t, I think you should fuck off now.	yes	no	yes	yes
Why do we allow Mexicans to work in our country!! Send them all back.	yes	yes	yes	no
F*ck them all jews!	yes	yes	yes	yes

Table 1: Examples of delicate texts compared to hate speech, offensive language, and profanity.

In this study, we target a broader set of sensitive texts that we call "delicate texts," an umbrella term covering toxic language as well as lower-severity sensitive texts, with a focus on sensitive texts (Table 1). Delicate text covers many topics which are not necessarily offensive but can still be highly sensitive and triggering. For example, texts where users share challenges regarding their mental health issues, where they discuss their experience of the loss of a loved one, or where they share content about self-harm and suicide. While most of these texts do not contain offensive language or attack certain minority groups, they all contain triggering topics that are emotionally and personally charged. Conversations about these topics can be easily derailed and lead to users experiencing discourteous or offensive behaviors from other users or virtual agents. With delicate text detection, our goal is

* The names of authors are arranged in reverse alphabetical order.

[1] Here, we use the terms "toxic language" and "hate speech" interchangeably.

[2] https://github.com/grammarly/detexd

[3] https://huggingface.co/grammarly/detexd-roberta-base

[4] https://huggingface.co/grammarly/detexd

The 7th Workshop on Online Abuse and Harms (WOAH), pages 14–28

to identify texts where engagement by other users or agents is most likely to result in harm, rather than focusing only on texts where harmful content has already been generated. Automatic detection of delicate texts is an essential tool for effective monitoring and prevention of potentially harmful content generated by users or AI. This model can be used for practical applications such as content moderation for models that are at high risk of hallucinations or data sampling to efficiently target texts where offensive interactions are most likely to happen.

In this study, we introduce the task of delicate text detection. We present a comprehensive definition of delicate texts and a dataset of 1,023 labeled delicate texts (DeTexD). We share our data collection, annotation, and quality control methods along with the detailed annotation schema. We describe the development of our baseline delicate text detection model. Finally, we demonstrate the difference between delicate text detection and existing content moderation methods by testing our model against toxic language benchmark datasets and testing popular content moderation models against our DeTexD dataset.

2 Related Works

Several studies have investigated the use of various NLP methods to detect inappropriate content; most of these works targeted toxic and offensive language. Some focused on developing more robust models to detect hateful content (Sohn and Lee, 2019), (Caselli et al., 2021), (Yousaf and Nawaz, 2022), while others focused on building better and less biased datasets (Founta et al., 2018), (Zampieri et al., 2019), (Basile et al., 2019), (Davidson et al., 2017), (Kiela et al., 2020), (Mathew et al., 2021), (Xia et al., 2020), (Huang et al., 2020), (Mollas et al., 2022), (Qian et al., 2019). With respect to dataset creation, (Mollas et al., 2022) created ETHOS, a binary and multi-labeled dataset of hate speech, along with a detailed annotation protocol. Their dataset covers various hate speech categories (including race, gender, religion, nationality, sexual orientation and disability), as well as target and whether the texts incited violence. They also examined the quality of their data using both binary and multi-label classification. In another study, (Mathew et al., 2021) created HateXplain, a hate speech dataset that reflects annotators' rationale for their labeling task. Their data went through a three-step annotation process in which a text was first classified as "offensive," "hate," or "normal;" next, the target of the hate was identified as "individual" or "generalized." Last, the annotators were asked to highlight parts of the text that justified their annotation decisions. They reported that models that used annotators' rationale in the training data performed slightly better than those without human rationale. Most studies have targeted hate speech; however, some studies have addressed a more general concept: inappropriate content. While most of these works used the term 'inappropriate content' to refer to hate speech, they also included sensitive topics. For instance, (Yenala et al., 2018) focused on identifying inappropriate content; they defined inappropriate content as impolite and disrespectful posts that offend certain groups, are related to illegal activities, or induce violence. They developed a deep learning-based model to identify inappropriate content in detecting query completion suggestions and user conversation texts. In another study, (Tripathi et al., 2019) focused on detecting sensitive content in user interactions with voice services. They targeted profanity, insult, geopolitical topics, explicit sexual and anatomical content, weapons, war, explicit graphical violence, race, religion, and gender. They focused on binary classification of sensitive content.

In (Basile et al., 2019) SemEval 2019 Task 5 dataset is described, a specific case of hate speech against immigrants and women in Spanish and English Twitter messages. They provide both: a main binary subtask for detecting the presence of hate speech, and a finer-grained one for identifying features such as hate attitude or target. During this competition, over 100 models were submitted. We evaluate our baseline model for delicate text detection on their main dataset.

3 Delicate text

3.1 Definition

We define *delicate text* as any text that is emotionally charged or potentially triggering such that engaging with it has the potential to result in harm. This broad term covers a range of sensitive texts that vary across four major dimensions: 1) riskiness, 2) explicitness, 3) topic, and 4) target. Delicate texts come with varying levels of risk; some can be highly risky such as texts about self-harm or content that promotes violence against certain identity groups, while others can be less risky such

as insulting language. Delicate texts can also have various degrees of explicitness: some can be produced explicitly with the use of delicate key terms, while others can be produced implicitly without the presence of delicate lexical terms. Delicate texts cover various subjects, with topics ranging from race, gender, and religion to mental health, socioeconomic status, or political affiliations.

Unlike toxic language that only targets identity groups (Zampieri et al., 2019), (Davidson et al., 2017), delicate texts can target identity groups, non-identity groups, or they can be self-targeted or non-targeted. In addition, delicate texts are not always offensive, unlike toxic language. Table 1 shows how different texts would be treated under our delicate text approach as compared to typical approaches for categorizing hate speech, offensive language, and texts containing profanity.

Table 2 illustrates examples of both delicate and non-delicate texts. The first "non-delicate" text does not contain any references to delicate topics. The rest all reference a delicate subject (mental health), but each has a different level of risk. For example, the "very low risk (1)" delicate text contains a factual statement about mental health, while the "very high risk (5)" text explicitly mentions self-harm. There is a shift in riskiness as content becomes more personal, emotional, and explicit. It is worth noting that none of these examples are offensive or contain vulgar language; however, engagement with these texts, whether by users or virtual agents, can result in harm.

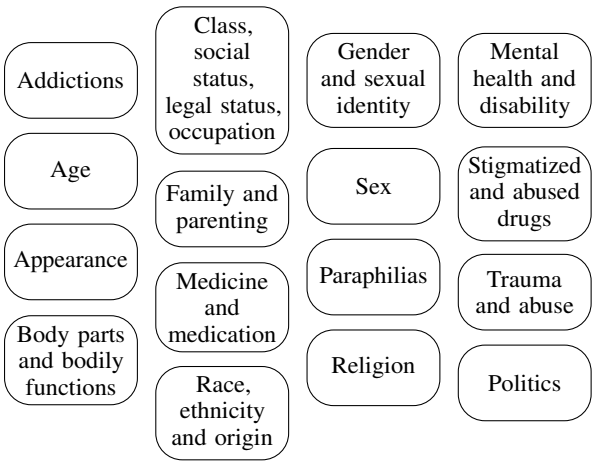

Figure 1: List of delicate text topics.

Figure 1 displays the list of delicate topics. We did not present this list in any hierarchical order as we wanted to highlight the fact that there is not a clear border between toxic language and sensitive language as a text can be both toxic and sensitive. Each of these topics can be used to create risky content. While these topics may be used in various degrees of riskiness, they all are considered delicate. Please see Appendix A for our annotation guidelines and additional examples.

4 DeTexD Benchmark Dataset

4.1 Data Collection

The sparsity of delicate texts in online platforms makes it challenging to target in data collection. To ensure that the data contained sufficient sensitive content, we used a combination of domain specification and keyword-matching when sourcing data. For our DeTexD Benchmark dataset (Table 3), we extracted data from various websites in CommonCrawl[5], where we specifically targeted news websites, forums discussing sensitive topics (e.g., Mental Health Forum[6], and able2know[7] which covers body image), and generally controversial forums (4Chan[8], Stormfront[9]), with the expectation that these would contain a significant amount of sensitive content. To further refine the dataset, we targeted paragraphs containing words from our delicate keyword dictionary. Our dictionary contains keywords related to a wide variety delicate topics which are included as tags in the dictionary metadata (see Table 6 for the full list of topics and examples of keywords). Additionally, each keyword is tagged with one of four severity ratings ranging from highly offensive to potentially offensive. We used these keywords with their metadata as a fine-grained data filtering method to extract delicate texts covering various topics and levels of risk. After analyzing the content of the pilot dataset we confirmed that the targeting methods resulted in a desirable distribution of data and moved forward with extracting 41,000 paragraphs through this method for the main annotation task. This data was split into DeTexD Train dataset containing 40,000 samples (this part is not published) and the DeTexD Benchmark dataset of 1,023 paragraphs. All data was collected from September 2019 to August 2022.

[5] https://commoncrawl.org/
[6] https://www.mentalhealthforum.net/
[7] https://able2know.org/
[8] https://www.4chan.org/
[9] https://www.stormfront.org/forum/

Text	Identification label (step I)	Riskiness label (step II)	Model prediction
A survey-based study of digital media usage shows options covering just about every topic, from true crime to mental health; there's certainly no shortage of options.	non-delicate	non-delicate (0)	NON-DELICATE
Identifying and addressing mental health, especially in the early stages, is highly important.	delicate	very low risk (1)	
Working with children who might have various mental health issues is such a challenging job and people need to understand how hard it can get at times.	delicate	low risk (2)	
One of my patients started having a type of mood swings, and he thought "this is normal for my age group" so he didn't seek help, he was unaware that all the males in his family had bipolar disorder, so the symptoms were left for about a year where they gradually got much worse and far more rapid. Ignorance is not bliss there is no harm in seeing your GP.	delicate	medium risk (3)	DELICATE
So you had a panic attack last night, and it seems that there was an issue with your new pills, and now you have an irrational fear alongside these pills. You imply that you've never had any major problems whilst out, so all I can suggest is to mention it to your GP next time you go.	delicate	high risk (4)	
I am just sick and tired of this life, and there is no hope for me, I am just looking for an easy way to end my life.	delicate	very high risk (5)	

Table 2: Examples of texts and corresponding labels from the DeTexD Benchmark dataset. "Identification" and "riskiness" are labeled by expert annotators (Section 4.2). "Model prediction" illustrates the mapping between 1-5 riskiness labels and binary NON-DELICATE and DELICATE predictions made by our baseline model (Section 5.1). Only predictions with a medium risk (3) or higher rating are converted to a DELICATE prediction.

4.2 Data Annotation

Identifying delicate content is a nontrivial task, as delicate text is a highly subjective concept. To ensure consistent and accurate annotations, we developed a fine-grained annotation scheme to guide expert annotators through the task. The annotation guidelines (Appendix A) contain our definition of delicate text along with a list of delicate categories within delicate text, examples of each category, and labeling samples.

To reduce the impact of subjectivity, we designed a two-step annotation scheme (Table 2):

Step I (identification): Annotators were shown texts and asked to label them as "non-delicate" or "delicate" based on our overall definition of delicate text. This initial binary rating pass allowed us to quickly identify texts most likely to contain delicate content through a relatively low-effort task.

Step II (risk level rating): Texts labeled "delicate" in Step I moved on to Step II, where annotators were asked to rate the risk level of each text on a riskiness scale of 1 ("very low risk") to 5 ("very high risk"). The annotators were instructed to focus on overall sentiment of the texts rather than the lexical meanings of individual keywords. Delicate texts which are more emotional, personal, charged, or those that reference a greater number of delicate topics are considered high risk, whereas texts with more neutral and less personal content are considered low risk (see Table 2 for examples of risk ratings).

Using this labeling process, we stepped away from simple binary labeling of the data, which not only helped to ensure quality, but also allowed us to gain more detailed information about the riskiness of the sensitive data.

Identification label (step I)	# samples	Riskiness label (step II)	# samples
non-delicate	503	non-delicate (0)	503
delicate	520	very low risk (1)	67
		low risk (2)	113
		medium risk (3)	153
		high risk (4)	113
		very high risk (5)	74
TOTAL (step I)	1023	TOTAL (step II)	1023

Table 3: Distribution of annotated texts in the DeTexD Benchmark dataset.

4.3 Quality Control

Each text in the DeTexD Benchmark dataset was annotated following the guidelines (Appendix A). All annotators who participated in this task are expert linguists that had an excellent understanding of delicate texts and had previously completed similar annotation tasks. Each text was annotated by three different annotators, and we took a majority vote as the final label.

To ensure annotation quality, we conducted a pilot annotation. Each snippet was annotated by one annotator, and 500/1,023 labeled snippets were randomly selected and qualitatively analyzed by the team of four expert linguists who designed the task. Each sample was reviewed, and its label was accepted if it matched the guidelines and rejected otherwise. Out of 500 judgments, 426 snippets were accepted, and only 74 labels were rejected (85% acceptance rate). After the pilot, the annotators were provided with feedback for improvement and the guidelines were updated to address common areas of confusion. Annotators who passed the pilot task moved on to annotate the full dataset. We measured the inter-rater agreement and obtained a Krippendorff's alpha score of 0.65 for the final dataset.

5 Experiments

In this section, we share results from a series of experiments. First, we show the potential to create a delicate text detection system which is suitable for practical usage. For this purpose, we developed and evaluated a baseline delicate text detection model. Next, we demonstrate the originality of this task by evaluating the performance of our model on toxic language datasets and evaluating toxic language detection models on DeTexD. Since toxic language detection and delicate text detection are two distinct tasks, DeTexD does not perform well on toxic language benchmarks and other content moderation methods that target mainly toxic language do not perform well on the DeTexD benchmark dataset.

5.1 Baseline Model

Our baseline model is the RoBERTa-based classifier (Liu et al., 2019b), which is fine-tuned on the delicate text detection training dataset of 40,000 samples.[10] The model is trained for 2,000 optimization updates on batches of 256 samples each. We used AdamW as an optimizer with a learning rate of $\alpha = 5e^{-5}$. As a task to learn, we selected a multiclass classification model with binary conversion because it has higher quality than binary classification and ordinal regression (Cheng, 2007). Although we noticed a better diagonal-aligned confusion matrix for ordinal regression, the evaluation result did not show a statistically significant improvement. In our settings, we train a 6-class classification model, where the classes are defined by the riskiness levels from annotation step II. The model's prediction is converted to a binary label using the mapping (Table 2):

i) *NON-DELICATE = non-delicate (0) ∪ very low risk (1) ∪ low risk (2)* and

ii) *DELICATE = medium risk (3) ∪ high risk (4) ∪ very high risk (5).*

5.2 Baseline Model Performance on Hate Speech Tasks

In order to experimentally confirm that delicate text detection and toxic language detection are distinct tasks, we ran our baseline delicate text detection model (Section 5.1) on popular toxic language datasets (Table 4).

[10] We are not publishing the training portion of our delicate text detection dataset, but it was annotated in exactly the same way as the DeTexD Benchmark dataset (Section 4).

Dataset	Model	Prec.	Rec.	F1
(Davidson et al., 2017), hate speech + offensive	(Davidson et al., 2017)	91%	90%	90%
	(Mozafari et al., 2020)	92%	**92%**	**92%**
	our baseline model	**95.2%**	70.5%	81.0%
(Davidson et al., 2017), hate speech only	(Davidson et al., 2017)	44%	61%	51%
	our baseline model	**60.9%**	**79.5%**	**69.0%**
(Founta et al., 2018)	our baseline model	76.3%	66.6%	71.1%
(Basile et al., 2019) SemEval-2019, Task 5A	(Basile et al., 2019)	**56.1%***	77.3%*	**65.0%***
	(Caselli et al., 2021)	48.3%	**96.4%**	64.5%
	our baseline model	47.5%	89.0%	62.0%
(Zampieri et al., 2019), OLID, Task A	(Zampieri et al., 2019)	**78%**	63%	70%
	(Liu et al., 2019a)	75.8%	**74.6%**	**75.2%**
	our baseline model	48.1%	66.4%	55.8%

Table 4: Performance of our baseline model on toxic language detection tasks as compared to the performance of models from the literature. *For the SemEval-2019 original model, only the accuracy and macro F-score were reported, so we inferred precision and recall values by numerically solving a system of equations with TP, FP, TN, and FN as unknown variables.

The Automated Hate Speech Detection (AHSD) dataset from (Davidson et al., 2017) has separate classes for offensive speech and hate speech, with examples labeled as hate speech representing the minority of the dataset (1,430 examples out of 24,783 total). We evaluate the performance of our model separately on the entire (Davidson et al., 2017) dataset, as well as only on the hate speech subset (which excludes all offensive speech examples). In both cases, after performing error analysis we can see that this dataset is not relevant for our classifier as the task is significantly different. Some false positive prediction examples, such as those mentioning race-related topics or explicitly sexual content, would be categorized as true positives in the DeTexD annotation schema although they are labeled as negative in this dataset. Most of the false negative prediction examples strongly correlate with specific offensive words such as "h*e" or "b*tch." Given the context in which these words are used, these examples would fall under the true negative definition of delicate text. Notably, our classification performance on the hate speech only subset exceeds that of the original work. We attribute the high performance to the fact that there is some overlap in the tasks specifically under the hate speech case, and that we use a more recent model architecture (Liu et al., 2019b), a pre-trained base model and larger model size.

After evaluating our baseline model on the dataset from (Founta et al., 2018) we found that it

is not relevant for our classifier as the task is very different from ours. A large proportion of false positives would be classified as delicate under our definition (e.g., sensitive topics such as "killing of thousands..."), while many false negatives would be classified as neutral according to our definition. However, here they are treated as overly emotional like "I'm fu***d up". After evaluating on SemEval-2019, Task 5, Subtask A (Basile et al., 2019) we found that it is not relevant for our classifier as the task is different from ours; it consists mostly of hate speech against migrants and women. As a result, false positives occur in instances where DeTexD detects other delicate topics, even including hate speech that is not targeted at women and migrants. False negatives occur in instances where refugee-directed hate speech is very specific and context-dependent such as "build that wall." Besides that, the dataset is unbalanced: for example, the word "b*tch" appears in half of the texts.

Offensive Language Identification Dataset (OLID), Subtask A (Zampieri et al., 2019) contains examples labeled as "offensive" or "not offensive." Similarly to our other evaluations, we find that the labels in this dataset do not significantly agree with our definition of delicate text. Many of the examples labeled as offensive in this dataset either do not contain enough context to make such a judgment (e.g. "A dying sport") under our definition, or look entirely neutral according to our definition of delicate text (e.g. "Yes. Yes he is!").

These experiments show that there is partial overlap between the definition of delicate text and commonly used definitions of offensive language and hate speech, which results in 70%-90% relative F-score of our baseline model for delicate text detection compared to models trained for toxic language detection (Table 4).

5.3 Comparing our baseline model and hate speech detection methods on the DeTexD Benchmark dataset

In order to evaluate our baseline model's performance and compare it with the most popular existing solutions for hate speech detection, we run them on our DeTexD Benchmark dataset (Table 5).

In our experiments, HateBERT models ("AbusEval", "HatEval", and "OffensEval") are the instances of HateBERT (Caselli et al., 2021), which are fine-tuned on the corresponding dataset. The highest precision is shown by the "HatEval" model,

Method	Prec.	Rec.	F1
HateBERT, AbusEval	86.7%	11.6%	20.5%
HateBERT, AbusEval#	57.0%	70.2%	62.9%
HateBERT, HatEval	**95.2**%	6.0%	11.2%
HateBERT, HatEval#	41.1%	**86.0**%	55.6%
HateBERT, OffensEval	75.4%	31.0%	43.9%
HateBERT, OffensEval#	60.1%	72.6%	65.8%
Google's Perspective API[11]	77.2%	29.2%	42.3%
OpenAI content filter[12]	55.0%	64.0%	58.9%
OpenAI moderation API[13]	91.3%	18.7%	31.1%
Our baseline model	81.4%	78.3%	**79.8**%

Table 5: Comparison of our baseline model for delicate text detection and existing hate speech detection methods on the DeTexD Benchmark dataset. HateBERT model here is from (Caselli et al., 2021).

which is fine-tuned on SemEval 2019 Task 5 dataset that contains hate speech against migrants and women (Basile et al., 2019). These topics are explicitly presented in the DeTexD dataset under "Gender" and "Nationality"/"Race" categories (Fig. 1). "OffensEval" shows the best overall performance among the HateBERT models. We speculate that this is because the definition of offensive language in the training dataset of this model (Basile et al., 2019) ("contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct") is broader compared to "AbusEval" and "HatEval," and it has greater overlap with our definition of delicate language. All HateBERT models show relatively low recall values because each HateBERT instance targets a narrow range of topics. After receiving valuable feedback from reviewers, we also calibrated optimal thresholds for f-score (marked with hash#). While F-scores got much higher, the precisions got much lower, so we consider this more like metric hacking. In real life, the proportion of positive cases is much lower, so for the future versions it may make sense to get test dataset with more negative cases.

Google's Perspective API is designed to moderate human interaction to support a friendly conversation environment. The Perspective API targets text attributes such as toxicity (rude, disrespectful, or unreasonable comments), severe toxicity (very hateful, aggressive, disrespectful comments), identity attacks (hateful comments targeting someone because of their identity), insults (insulting comments toward people), profanity (swear words), and threat (intention to inflict pain, injury, or violence against people)[14]. This target definition is similar to the "OffensEval" dataset, which could explain why performance is similar to the HatEval model.

[14] https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages

The OpenAI content filter shows strong recall but is lacking precision in our experiments. In error analysis, we see that it misses a large part of examples from mental health and medical topics. In a sample of false-positive predictions, we only see a slight pattern of a tendency to flag texts that contain profane keywords. Surprisingly, during our testing of the OpenAI content filter, we found that for about half of the inputs the predictions are stochastic, with standard deviation on binary prediction reaching as high as 0.5 (across 100 predictions). We expect the presented results for the OpenAI content filter to have a wider than expected confidence interval.

The authors of the OpenAI moderation API suggest it as a replacement for the OpenAI content filter. On our benchmark dataset, the moderation API has higher precision but lower recall as compared to the OpenAI content filter. This can be explained by the difference in the definition of target content between the two models. During error analysis, we find that lower recall can mostly be attributed to medical and mental health topics in our dataset, although some of the examples relating to sexual content were also missed. All examples where the OpenAI moderation API made a false-positive prediction relate to sexual content or socioeconomic status categories. However, the sample is too small (6 out of 687 non-delicate) to make strong conclusions.

In summary, our experiments show that none of the studied toxic language detection methods provide satisfactory detection performance in delicate text detection. Most commonly, the evaluated hate speech detection methods either miss coverage on medical and mental health topics, show lower precision on examples that contain offensive keywords (but aren't deemed delicate according to our definition), or both.

6 Conclusions

We introduced a new type of sensitive language called "delicate text," an umbrella term covering not only toxic language but also sensitive language with a priority focus on the latter. We annotated the DeTexD Benchmark dataset for delicate text detection. The significance of the difference in the definitions is highlighted by the relative performance deltas between models trained each definitions and corpora and evaluated on the other. We make our annotation guidelines, annotated dataset, and baseline model publicly available.

7 Limitations

The presented results only apply to the English language. Both our benchmark dataset and the baseline model target the English language exclusively. Special text sources such as instant messaging or speech-to-text are likely under-represented in our benchmark test set; therefore, we did not evaluate classification performance in those domains. Since we used RoBERTa as the base model, our model inherits the same limitations. Specifically, the length of input sequences is limited to 512 BPE tokens, and additional pre- and post-processing is necessary to run predictions on longer inputs. However, we did not evaluate prediction aggregation methods or classification performance.

7.1 Gender bias

We expect our baseline model to exhibit some amount of gender bias, since it is based on a model that was trained on a wide variety of human-written text. However, in our analysis we only found limited evidence of gender bias. In the counterfactual prediction difference analysis on the DeTexD dataset, we found 6 examples (0.6% of 1023 total) for which the baseline model made a different prediction depending on gender context. The counterfactual texts were created by performing binary gender-swapping (male/female) of the nouns, pronouns and gendered names.

In gender bias evaluation using the Winogender Schemas dataset (Rudinger et al., 2018), we found 4 example triplets (1.7% of 240 total) where the baseline model made different predictions depending on the gender context. We noticed a slight correlation between gender and our model's binary prediction, but consider it negligible at this sample size.

Due to the limited volume of evidence provided by our experiments we cannot make strong conclusions about the presence or absence of gender bias in our baseline model. However, our understanding is that for the best results, bias mitigation methods must be applied. One such method could be the gender-swapping data augmentation of the training data.

7.2 Country bias

The baseline model described in this paper (Section 5.1) treats different countries differently. We

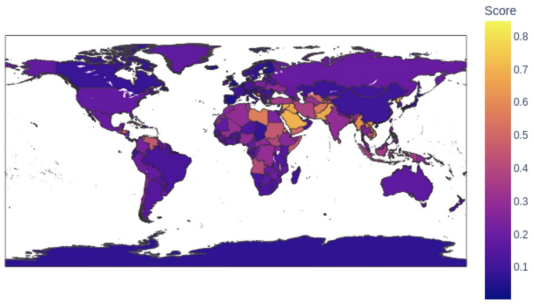

Figure 2: Illustration of country bias.

measure this effect by calculating scores for the template phrase "Unfortunately, the incident happened in {country_name}." The result is illustrated on the colored map on Figure 2. One way to fix this is to replace country names with random others in training data. However, in this case some sentences can become nonsense. For example, "London is the capital of the United Kingdom." In this case, we would need to change all dependent words which would become too complex. From our observation, this situation happens with half of all sentences including country information, so we leave this part as an open question for now.

Ethics Statement

Hate speech, offensive language, and delicate texts are sensitive, and very important matters. Through this work, we try to dive deeper into the challenges and opportunities of any delicate text detection. The goal of this work is to expose the strengths and limitations of different delicate text detection and related techniques and their implications. Some datasets, and models that we work with have been publicly released for a couple of years. All of these artifacts are considered to be in the public sphere from a hate speech perspective. We do not make any recommendations on using these on public or private datasets without proper due diligence for privacy, security, sensitivity, legal, and compliance measures.

Please be advised that due to the nature of the subject matter, the presented DeTexD Benchmark dataset includes a variety of uncensored sensitive content, such as hate speech, violence, threat, self-harm, mental health, sexual, profanity, and others. The text of this work includes keywords and partial text examples of the same type. The most extreme occurrences of such examples in this text are partially obscured with asterisks but the semantics are retained.

8 Acknowledgements

We express our gratitude to our colleagues Cortney Napoles and Leonardo Neves for their valuable advice and to our managers Viktor Zamaruiev and Max Gubin for their constant support. To our communities: While we are writing this, our homeland Ukraine continues to resist the unprovoked Russian invasion. We are grateful to everyone who defends Ukraine, declares support to the people of Ukraine, and is sending aid. Thank you!

References

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Jianlin Cheng. 2007. A neural network approach to ordinal regression. *CoRR*, abs/0704.1028.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018b. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael

Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Rafal Kocielnik, Shrimai Prabhumoye, Vivian Zhang, R Michael Alvarez, and Anima Anandkumar. 2023. Autobiastest: Controllable sentence generation for automated and open-ended social bias testing in language models. *arXiv preprint arXiv:2302.07371*.

Ping Liu, Wen Li, and Liang Zou. 2019a. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Andrew C Parnell, Víctor González-Castro, Rocío Alaiz-Rodríguez, and Gonzalo Molpeceres Barrientos. 2020. Machine learning techniques for the detection of inappropriate erotic content in text. *International Journal of Computational Intelligence Systems*, 13(1):591.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018a. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018b. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. 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.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Rahul Tripathi, Balaji Dhamodharaswamy, Srinivasan Jagannathan, and Abhishek Nandi. 2019. Detecting sensitive content in spoken language. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 374–381. IEEE.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech

detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Harish Yenala, Ashish Jhanwar, Manoj K Chinnakotla, and Jay Goyal. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6:273–286.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. 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.

A Appendix. Guidelines for building DeTexD.

A.1 Glossary

In these guidelines, you are going to see numerous references to sensitivity in language as well as related notions. Before reading the document, please familiarize yourself with the following terms that will help you get a better understanding of the task:

- **Delicate** (adj. for a text/word/subject matter): referencing a touchy or sensitive subject. This includes texts that are emotionally charged and that cover topics that are potentially triggering, offensive, taboo, intimate, or about marginalized groups.
- **Delicate topics**: topics that are usually delicate. Examples include mental and physical health-related topics, trauma and violence, or identity-related topics. See an the list of sensitive topics in Table 6.
- **Delicate keywords**: words that semantically relate to a certain delicate topic. For example:
  - *democrat, chauvinist, islamo-leftism* are normally used in political language,
  - *able-bodied, autistic, bulimia* will typically refer to the topic of ableism or mental health.

A.2 Delicate topics

Table 6 provides the list of delicate topics and the definitions of their typical language that can be associated with the language of a certain delicate topic.

A.3 Context

This task will ask you to make two judgment steps:

Step I. Identification: decide if a text is delicate or non-delicate using the following definitions:

- A **delicate** sentence contains emotionally-charged references to a sensitive topic.
- A **non-delicate** is fully innocuous and doesn't contain any particularly charged references to a sensitive topic.

Step II. Riskiness estimation: rate how **delicate** texts are using a 5-point scale where 1 stands for "very low risk" and 5 for "very high risk". How delicate a sentence is must be evaluated with regard to its sentiment rather than the lexical meanings of separate keywords. In other words, the more emotional and personal the tone, the more delicate the sentence. The following questions can help you make a decision:

(a) Is the content of the sentence emotionally charged rather than factual?
(b) Can the content of the sentence evoke negative feelings?
(c) Does the content of the sentence pertain to a sensitive topic and show bias against particular groups of people?

If the answer to any of these questions is positive, the sentence will fall on the high end of the riskiness scale. Find a detailed interpretation of the riskiness scale in the next section.

See the examples of annotated texts in Table 8.

A.4 Riskiness estimation

Table 9 provides a description of riskinesses that are likely to fall on certain parts of the rating scale, examples of delicate texts, and explanations.

A.5 Paragraph-level judgments

The paragraphs are annotated holistically. This means that the assigned score is not based on just the individual sentences within a paragraph, but rather the score is reflective of the overall meaning of the paragraph. However, the score can be affected by the number of:

- delicate sentences within a paragraph;
- explicit delicate sentences within a paragraph;
- emotional and personal sentences within a paragraph;
- paragraphs that have a higher number of delicate sentences, higher level of explicitness, and have more emotional and personal weights should get a higher score. A comparative analysis of some examples is presented in the Table 10.

Delicate topic	Description	Examples of related delicate keywords
Addictions	Language associated with addictive behavior.	alcohol, gambling, toxicomania
Age	Language associated with biological age, age identity, and age discrimination.	elderly, elderspeak
Appearance	Language that has to do with physical appearance and prejudice based on physical appearance.	unibrow, humpback
Body parts and bodily functions	Language used to talk about sensitive body parts and bodily functions.	breast, wiener
Class, social status, legal status, occupation	Language used to talk about people in the context of their economic, legal, and cultural factors.	yuppie, unserved, refugee
Crime and violence	Language describing violence, crime, and people who are engaged in it.	murder, arsonist, genocide.
Family and parenting	Language associated with marital status, parental status, adoption.	co-parenting, surrogacy
Gender and sexual identity	Language associated with the lgbtqai+ community or sexual orientation.	heterosexual, agenderfluid, cross-dresser
Medicine and medication	Language associated with diseases, illnesses, and medication.	venereal, abortion, chancre
Mental health and disability	Language used to talk about people and issues related to disability and mental health.	able-bodied, autist, bulimia
Paraphilias	Language related to perversions.	asphyxiophile, frotteurism
Politics	Language associated with political ideas and movements as well as their representatives, advocates, and opponents.	democrat, chauvinist, islamo-leftism
Race, ethnicity, and origin	Language that has to do with racial and ethnic identity.	white, brown
Religion	Language associated with religious ideas and movements as well as their representatives, advocates, and opponents.	bible, burqa, mosque
Sex	Language describing sex-related activities and their participants, sex toys.	cowgirling, penetrate, dildo
Stigmatized and abused drugs	Language used to talk about stigmatized or abused medications.	viagra, clonazepam, barbiturate
Trauma and abuse	Language referencing physical as well as psychological trauma and abuse.	victim, bullying, gaslighting

Table 6: List of delicate topics with descriptions and delicate keywords.

Delicate topics	Count
Hate speech, race, religion, sexual orientation, nationality	45
Mental health, physical and mental issues, disability	166
Socioeconomic status, political affiliation, fake news, misinformation	17
Self-harm	17
Sexual content, pornographic and erotica	81
Crime and violence, threat and incitement to violence	36
Other	30

Table 7: Distribution of delicate topics.

Text	Identification label (step I)	Riskiness label (step II)	Explanation
Both women suspect that Lars has his wife's gold wedding band in the handbag.	non-delicate	non-delicate (0)	No references to delicate topics.
Huh I guess that explains why Zelina's cat is always on Black's cats' Instagram page.	non-delicate	non-delicate (0)	No references to delicate topics.
Ongoing research projects are focused on quantitative methods for evaluating animal systems (production, reproduction, mammary health; the value of milk urea nitrogen (MUN) as a nutrition management tool; feed cost optimization and nutritional economics; and production risks and risk management for dairy farms.	non-delicate	non-delicate (0)	No references to delicate topics
If someone rejects one and believes the other, they are not following science, they are following their personal ideologies.	non-delicate	non-delicate (0)	No references to delicate topics.
On the other hand, some women would question me about how much I enjoyed it and if I were ambitious in it.	non-delicate	non-delicate (0)	No references to delicate topics.
Everyone seems to have an agenda or view point, so it's rarely true science.	non-delicate	non-delicate (0)	No references to delicate topics.
I'd say, "Oh, that seems incorrect to me."	non-delicate	non-delicate (0)	No references to delicate topics.
I know things I've seen talked about in the past are, moving during his ult, or crippling people in his ult.	non-delicate	non-delicate (0)	The sentence includes "crippling" but is low risk.
There is a good expression, "kill with eyes," and I don't mind being killed with their eyes — with their blue, and ever blue eyes.	non-delicate	non-delicate (0)	Just a metaphorical expression.
I wonder if they're stream snipers.	non-delicate	non-delicate (0)	Obvious references to video games.
There isn't a cure for loss.	delicate	very low risk (1)	The sentence contains a subtle reference to a delicate topic, but it's rather an impersonal piece of wisdom.
I found out later they didn't throw it away but packed it up as punishment.	delicate	very low risk (1)	The word "punishment" refers to violence, but the sentence can be subject to multiple interpretations.
Why, am I under arrest?	delicate	very low risk (1)	The sentence may refer to a delicate topic, but 1) the sentence may be interpreted in a number of ways; 2) interrogatory sentences seem to be less risky than statements.
Not really it is the way humans dispose of plastic that is bad although I will say the materials used to make it suck but we use them anyway right now for other things.	delicate	very low risk (1)	The topic of ecology is relatively delicate, and the word "suck" makes the sentence slightly charged. However, the sentence doesn't seem to be particularly risky.
As for OP's question, I think you should come out.	delicate	low risk (2)	The sentence is subject to multiple interpretations.
But they're not outright shit.	delicate	low risk (2)	The sentence is a low-key insult that could apply to anyone. Negation lowers the delicateness.
I am a heterosexual female, married to a heterosexual male.	delicate	low risk (2)	The sentence introduces some facts connected with the author's sexuality that is unlikely to be associated with a stigma.
Women are particularly susceptible to developing depression and anxiety disorders in response to stress compared to men.	delicate	delicate (3)	The sentence is referencing mental health, but the tone is impersonal and academic.
Mumbai, October 17: Rahul Mahajan, the son of former BJP general secretary Pramod Mahajan would sit on a dharna and address a meeting to support the death sentence awarded to Parliament mastermind Mohammed Afzal Guru.	delicate	medium (3)	The sentence covers several delicate topics and refers to a particular person, but it should be counted as moderately risky since it's a news report.
Most importantly, powerless communist labor will always be cheaper than democratic labor because voting and civic education are expensive.	delicate	medium (3)	The sentence is dry and talks about things in the abstract.
The supported employees ranged in age from early twenties to late forties, none were from ethnic minorities and only four of them were women, reflecting the demographics of supported employees in the area.	delicate	medium (3)	The sentence is quite neutral. The density of sensitive references makes the sentence potentially problematic, but given that the sentence is impersonal and not emotional, it doesn't deserve a (4) or (5).
I used to live in that district next to the bank and right across from where that guy got shot last year.	delicate	high risk (4)	The sentence introduces a tragic event.
So people who disapprove of prenatal testing, abortion or reproductive technologies should be under no obligation to undergo any such treatments themselves.	delicate	high risk (4)	The sentence is likely to be presenting the author's attitude towards a delicate topic.
I started to get a little freaked out, thinking maybe a stranger was in the house, but all of a sudden my sister yelled 'Elf', her service dog's name and said there was blood on his tail.	delicate	high risk (4)	Personal experience associated with fear and possibly violence.
I get the euphoria side effects of codeine so I tend to drown myself in that and Xanax on days I know I'll be spending in bed.	delicate	very high risk (5)	The sentence is very personal and emotional. The reference to a delicate topic is undeniable.
This dbag also banged a lot of females that he worked with.	delicate	very high risk (5)	The sentence contains offensive words and introduces face-threatening facts about certain people (even though their names are not disclosed).
wait what happened to your sisters mental health exactly and at what age?	delicate	very high risk (5)	The sentence is mentioning a concrete delicate story associated with close people.
OP, I know it's different dating someone with PTSD especially if you never have or if you don't have it yourself.	delicate	very high risk (5)	A subjective statement about a delicate topic.
There, Elster had plans to do away with his wife under circumstances that falsely implied that she had committed suicide.	delicate	very high risk (5)	Explicit crime-related content.
Difficult to consume enough to get high because of the nausea the drug produces.	delicate	very high risk (5)	The sentence is clearly referencing a delicate topic.

Table 8: Examples of texts which passed two annotation steps: identification (step I) and riskiness estimation (step II). In step I, the annotators were shown a text and were asked first to label the text as "non-delicate" or "delicate." In step II, if the annotators labeled a text as "delicate" in step I, they were asked to estimate the riskiness of the delicate text using a 5-point scale where (1) stands for "very low risk" and (5) for "very high risk."

Delicate text	Riskiness label (step II)	Motivation of riskiness label	Explanation
This is **fucking** amazing!	1	The sentence contains a sensitive keyword, but it doesn't engage with any delicate topics.	The f-word is used as an intensifier of the positive meaning.
God never received a PhD because he had only one major publication. If **Jesus** had had a car, it would have been a Christler.	1	God or Jesus are sensitive keywords that belong to the topic of religion. However, the content of the sentences is inoffensive and does not perpetuate stereotypes. The sentiment is well-meaning and kindly humorous.	The sentence refers to a delicate topic but the sentiment is humorous, and it does not perpetuate stereotypes.
I am a heterosexual female, married to a heterosexual male.	2	The sentence introduces some facts connected with the author's sexuality that is unlikely to be associated with a stigma.	The sentence contains sensitive keywords, but the sentence is not emotionally charged.
Assuming you use lube and don't try to force one that won't slide in.	2	The sentence doesn't allude to the topic of sex, but it may also have other non-sensitive interpretations.	The sentence may or may not relate to a delicate topic since it is subject to multiple interpretations.
30% of the people living in that country have depression. Vaginal cancer accounts for approximately three percent of cancers of the female reproductive system. These summits provide communities with opportunities for strategic planning and collaboration building to combat methamphetamine problems faced in their own communities.	3	Even though the sentences touch upon delicate topics, they present plain facts in an unemotional matter-of-fact way.	The sentences contain references to delicate topics, but its tone is neutral, academic, medical, or legal. It does not perpetuate stereotypes or bias.
Funny how the **conservatives** crying for cutting spending suddenly backtrack when it comes to pet projects they like, right? It was a blast watching my parents and all their friends lose their retirement savings under **Bush** and my cousin go off to war and come back having no clue what they accomplished.	4	The sentences contain notes of sarcasm and ridicule targeted at certain political representatives.	The sentence contains notes of negative sentiment or sarcasm targeted at certain political party.
I've been working on recovery for about 5 years for complex PTSD and was doing ok, but then 'bam' I was in severe physical pain, which amplified all of the emotional pain... I've been there every stage of my handsome man's transition and only get more and more attracted. Afterwards, I was transferred to an orphanage in Temuco called Hogar Mamá Augusta.	5	The examples denote the author's personal sensitive experience.	The sentence contains references to delicate topics, and the author is emotionally involved in it.
Ernest is having urges to use opioids and said you are supporting him with the issue. Years later my friend told me he had survivor's guilt and asked himself why he survived.	5	The sentences tell the story of certain people's sensitive experience.	The sentence narrates a tragic or emotionally charged story of a certain individual or a group of people (e.g., suicide note, doctor's notes, descriptions of abuse, etc.).
MDMA and LSD is an amazing drug combo and it will not do you any damage. Also, get the hell off Atripla, it causes psych issues and is no longer a tier 1 recommended medication.	5	Both sentences are about the author's personal perspective as they explicate strong opinions on delicate matters and offer advice.	The sentence expresses the author's strong stance on a delicate topic.
Fuck you! I will come after you and kill you.	5	The sentence contains a threat.	The sentence contains threats.

Table 9: Examples of delicate texts with explanations for various riskinesses. Note that this list is not exhaustive. If you encounter a text that doesn't match any of the descriptions, make the decision using the questions provided in Section A.3.

Text	Score	Explanation
I just hate talking about my mental issues with people; this makes me feel very uncomfortable. I sometimes think that if I share them with anyone, they will judge me, and I just want to not exist in a situation where people judge me.	5	Delicate sentences within a paragraph. All of the sentences in this paragraph are highly delicate and contain extreme emotional and personal content.
I hate talking about my mental issues with people. However, I am aware of the fact that mental issues are like other issues that need to be addressed and adequately taken care of. People must understand the significance of any mental illness and be able to react appropriately to these issues.	3	Delicate sentences within a paragraph. This paragraph contains one delicate sentence that expresses personal emotion, while the remaining sentences have delicate topics that are more factual.
I f***king hate you and people like you, you ugly piece of s**t. I f***king want to shoot you in the head!	5	Explicit delicate sentences within a paragraph. This is a highly delicate paragraph as it contains threats.
I don't feel comfortable around you and I prefer to not hang out with you. You have a very negative energy.	2	Explicit delicate sentences within a paragraph. This sentence does not have any explicit references to a delicate topic; however, it contains personal/emotional content.
The pain and suffering are so much, and I can barely endure it. I feel that I am being suffocated, and I don't want to live anymore.	5	Emotional and personal sentences within a paragraph. This paragraph is highly delicate with extreme emotional and personal content.
Some trauma can have long-lasting effects; the pain and suffering can become unbearable to the point that the patient might feel suicidal.	3	Emotional and personal sentences within a paragraph. This is an example of a paragraph that contains a very delicate topic; however, the topic is presented through factual statements.

Table 10: Paragraph-level annotations.

Question	Answer
Do definite referents make delicate sentences more sensitive than indefinite ones? E.g., *But they're not outright shit.* vs *But [some particular group] are not outright shit*.	Introducing a definite referent can increase the sensitivity of the sentence. However, it's unlikely to turn a non-delicate sentence into a delicate one. E.g., both *No doubt "nobody" would take the job if he was offered a decent pay* and *No doubt 'Tom the Nobody' would take the job if he was offered a decent pay* are non-delicate.
How should we treat mild second-person insults? E.g. *"You loon!"*, *"Your breath doesn't smell great."*	Delicate but low-sensitivity.
What kind of sentences should we consider to be incomprehensible and discard?	By incomprehensible, we mean anything that doesn't make sense at all. E.g.,*no iea why it wentjlout.*, *USFreighways jodohku*. Fragments like *both manipulative assholes lol* should be judged.
How delicate are news reports? E.g., *Mumbai, October 17: Rahul Mahajan, the son of former BJP general secretary Pramod Mahajan would sit on a dharna and address a meeting to support the death sentence awarded to Parliament attack mastermind Mohammed Afzal Guru*.	We assume that news reports are as delicate as academic/legal/etc. texts.
How does the density of delicate references impact the sensitivity level?	Sentences referencing multiple topics are likely to be more delicate than those referencing only one.
Sentences may be targeted at various addresses. How does that correlate with delicateness?	Pragmatic proximity and definiteness of an addressee escalate the sentence sensitivity. E.g., *I hate you.* would be more problematic than *I hate anyone who does things like that*.
Are sentences containing words like "father," "mom," "etc." is delicate since they refer to the domain of family and parenting?	It's contingent on the context: I went to the theater with *my mon yesterday* is non-delicate. *I'd like to find my biological dad* and *I'm his biological dad. are definitely delicate*.
Are sentences mentioning names associated with historical/religious figures delicate? E.g. *Good morning, Adolph*.	No. "Adolph" can be anyone else's name. However, references to concrete figures (Trump, Obama, Hilary(clearly used in a political context)) make the sentence delicate.
Are passages from religious texts delicate? Are these in-line with academic texts/news articles? E.g., *The Jews did not believe that he had been blind and had received his sight, until they called the parents of the man who had received his sight [21] But how he now sees we do not know, nor do we know who opened his eyes*.	They are delicate, but borderline ("very low risk (1)"). These are in line with academic/legal texts.
Are sentences referencing violence in video games delicate?	Please mark such sentences as non-delicate as long as you find explicit references to video games. *In other words, I've died 3 back to back games on 4 kills*. would be non-delicate as its connection with video games is quite obvious. Other cases (e.g., *I killed him with a sword*) would be delicate.

Table 11: Frequently asked questions about delicate text annotation.