Rethinking Offensive Text Detection as a Multi-Hop Reasoning Problem

Anonymous ACL submission

Abstract

We introduce the task of implicit offensive text detection in dialogues, where a statement may have either an offensive or non-offensive interpretation, depending on the listener and context. We argue that reasoning is crucial for understanding this broader class of offensive utterances, and create Mh-RIOT (**M**ulti-hop **R**easoning **I**mplicitly **O**ffensive **T**ext Dataset), to support research on this task. Experiments using the dataset show that state-of-the-art methods of offense detection perform poorly when asked to detect implicitly offensive statements, achieving only $\sim$0.11 accuracy.

In contrast to existing offensive text detection datasets, Mh-RIOT features human-annotated chains of reasoning which describe the mental process by which an offensive interpretation can be reached from each ambiguous statement. We explore the potential for a multi-hop reasoning approach by utilizing existing entailment models to score the transitions of these chains, and show that even naive reasoning models can result in improved performance in most situations. Analysis of the chains provides insight into the human interpretation process and emphasizes the importance of incorporating additional commonsense knowledge.

1 Introduction

With the development and popularity of online forums and social media platforms, the world is becoming an increasingly connected place to share information and opinions. However, the benefit these platforms provide to society is often marred by the creation of an unprecedented amount of bullying, hate, and other abusive speech[1]. Such toxic speech has detrimental effects on online communities, and can cause great personal harm. Some efforts by the NLP community to address this problem have achieved high accuracies in classifying toxic speech in specific domains, such as sexist (Golbeck et al., 2017), racist (Waseem, 2016), or otherwise hateful text (Ross et al., 2016; Gao and Huang, 2017; Davidson et al., 2017).

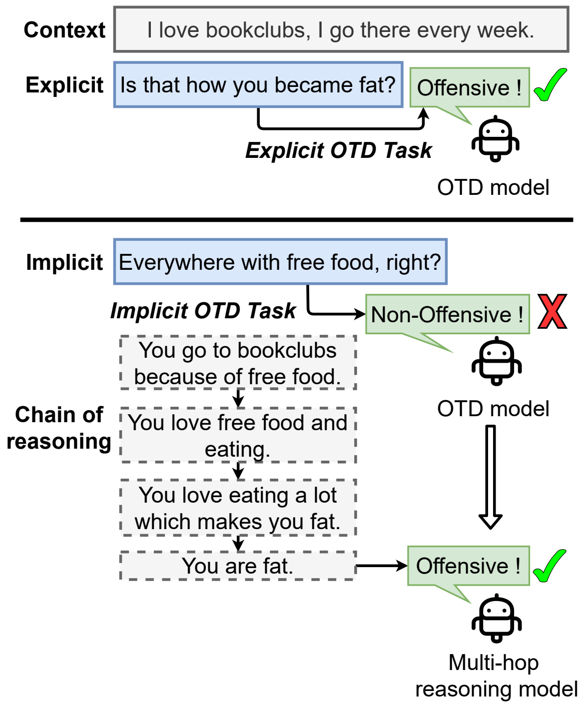

Figure 1: An instance illustrating Explicit OTD, Implicit OTD and our multi-hop reasoning approach.

While many instances of toxic speech are blatant and easily identified with sentence-level classifiers, not all offensive text contains obvious indicators. Waseem et al. (2017) argues for the classification of offensive text into two categories, (1) **explicit offensive text**[2], which is unambiguous in its potential to be offensive and often includes overtly offensive terms, such as slurs, and (2) **implicit offensive text**, which is more ambiguous, and may use sarcasm, innuendo, or other rhetorical

[1] Disclaimer: due to the nature of this work, data and examples may contain content which is offensive to the reader.

[2] Waseem et al.(2017) originally defined these terms as "explicit/implicit abusive text", but we adopt the phrase "offensive text" as used by the OTD community.

devices to hide the intended nature of the statement. In this work we argue that there exists a direct relationship between these tasks, and that each implicitly offensive statement corresponds to an explicitly offensive statement which is realized through the interpretation process. This explicitly offensive statement is closer to the sentiment the listener feels when interpreting the statement as offensive. Consider the example in Figure 1, a dialogue between two speakers, S1 and S2:

S1: "I love bookclubs, I go every week"

S2: "Everywhere with free food, right?"

By itself, the statement by S2 is innocuous and could be interpreted as a simple prompt for more information about the bookclub. However, other interpretations of this statement could lead S1 to arrive at a number of explicitly offensive statements, such as (1) "*You are poor*", (2) "*You are fat*", (3) "*You are not smart/sophisticated*". Thus we consider the chain of reasoning which constitutes the interpretation to be a crucial part of recognizing implicitly offensive statements.

The importance of more complex reasoning when resolving such ambiguities in offensive content is not new. The Hateful Memes dataset (Kiela et al., 2021) pairs images with unrelated text captions. Both of these components are benign when considered independently, but combining them can occasionally create memes with offensive interpretations. Consequently, approaches which jointly reason over a combined representations of each modality outperform those which treat each modality independently, hindering the system's ability to perform more complex reasoning.

To study this phenomenon purely in the text domain, we use human annotators to construct a dataset consisting of (1) an implicitly offensive statement, (2) a corresponding explicitly offensive statement, and (3) a chain of reasoning mapping (1) to (2). When evaluated on the explicitly offensive examples, state-of-the-art models perform well, achieving $> 90\%$ accuracy. However, when applied to the implicit OTD samples, the accuracy of the models drops to an average of about $< 11\%$. We then explore the use of a multi-hop reasoning-based approach by utilizing a pre-trained entailment model to score the transitions along each "hop" of the reasoning chain. When incorporating additional knowledge (from human annotations) into the premises of each entailment, we achieve higher accuracy than comparable methods which do not utilize the reasoning chain. We present this as evidence that a multi-hop reasoning-based approach is a promising solution to this problem, and release our data to support further research into this problem.

Our contributions in this work are threefold:

- We propose the task of implicit offensive text detection (Implicit OTD), and construct a dataset to research on this topic. The dataset contains annotations of reasoning chains to support study into multi-hop approaches.
- We conduct experiments using existing state-of-the-art OTD models, and show they perform poorly on Implicit OTD task.
- We examine the use of entailment models as part of a multi-hop reasoning approach for Implicit OTD, showing improved accuracy in most cases. We provide an analysis of which types of reasoning are most challenging, and which types of external knowledge is required.

2 Related Works

OTD in Text Classification Early approaches to OTD relied primarily upon dictionaries like hatebase [3] to lookup offensive words and phrases. The creation of OTD datasets enabled the development of ML-based approaches utilizing simple features, such as bag-of-word representations (Davidson et al., 2017). With the advent of social media platforms, many resources have been developed for identifying toxic comments in web text (Waseem and Hovy, 2016; Davidson et al., 2017), including a number of deep learning-based methods (Pitsilis et al., 2018; Zhang et al., 2018b; Casula et al., 2020; Yasaswini et al., 2021; Djandji et al., 2020). Notably, all of these methods can be described as building a contextual representation of a sentence (whether trained end-to-end or on top of existing pre-trained language models), and making a classification based on this representation.

OTD in Dialogue Systems As user-facing technologies, preventing dialogue systems from producing offensive statements is crucial for their role in society. As noted in Dinan et al. (2020), toxicity in generated dialogue may begin with biases and

[3] www.hatebase.org

offensive content in the training data, and debiasing techniques focused on gender can reduce the amount of sexist comments generated by the resulting system. Similar outcomes can be obtained through adjustments to the model or training procedure, for instance, toxic words can be masked during training to reduce their role in model predictions (Dale et al., 2021). GeDi (Krause et al., 2020) proposed using class-conditional LMs as discriminators to reduce the toxicity produced by large pre-trained LMs (GPT-2). Additionally it may also be important to identify offensive statements made *to* a dialogue system, as it has been shown that dialogue systems can react with counter-aggression (Cercas Curry and Rieser, 2018), and systems which continuously learn during deployment may incorporate toxic user responses into future generations.

Subjectivity in OTD Previous work has hit upon the role that an individual's own perspective may play when determining offensiveness. For instance, in the Offensive Language Identification Dataset (OLID), a widely used OTD dataset (Zampieri et al., 2019a,b, 2020), annotations exist on a hierarchy. Each level dictates the target of the offensive text, in terms of their identity as a group, individual, or entity. But to our knowledge, a person's identity or attributes have not played a critical role in existing OTD research. OLID was also augmented with labels for capturing the degree of explicitness (Caselli et al., 2020)), and may also support research into resolving implicitly offensive statements. However, implicitness in OLID is defined primarily as the lack of an overtly offensive word or slur, and the aforementioned personal attributes or subjectivity of interpretation are not considered. Our dataset differs in this respect, as we consider not just if a statement is offensive, but *how* it can be considered offensive, by defining the interpretation process as a chain of reasoning towards a subjective experience. In this sense, a more similar approach comes from normative reasoning in moral stories (Emelin et al., 2020), where a short chain of reasoning is used to assess morality of actions and consequences.

3 Data

We propose Mh-RIOT as a dataset for the study of Implicit OTD as a multi-hop reasoning problem, and for use as a diagnostic to test models' ability to identify implicitly offensive statements.

Each example in the dataset consists of three parts:

1. A personal attribute of the reader/listener.
2. An implicitly offensive statement, its corresponding explicitly offensive statement, and a non-offensive statement.
3. A chain of reasoning, describing the iterative process of how the ambiguity of the implicitly offensive statement can be resolved into the corresponding explicitly offensive statement. Appendix A lists some sample chains in Mh-RIOT.

We collect annotations for Mh-RIOT using Amazon Mechanical Turk (AMT). Four pilot experiments were conducted to select qualified annotators for the final annotation. The instructions provided to the annotators can be found in Appendix C.

3.1 Annotation Scheme

Personal Attribute As we have defined in Section 1, we argue that the context in which a statement occurs is crucial to understanding its potential in creating an offensive interpretation, and therefore the context should play an important role in the annotation task. However, providing an overly specific context can increase the difficulty of providing a relevant implicitly offensive statement. To make the annotation task more feasible we reduce the context to a single feature: a personal attribute of the reader/listener.

The set of attributes is obtained from the personas in the PERSON-CHAT corpus (Zhang et al., 2018a), of the form "*I like sweets.*", or "*I work as a stand up comedian.*" Attributes related to ethnicity, gender, and other protected classes are manually removed, leaving 5334 distinct attributes. We divide the attributes into several categories (detailed category information can be found in Appendix B) before randomly sampling a subset of 920 attributes, uniformly across categories, in order to increase the number of workers assigned to each attribute.

Implicit, Explicit and Non-offensive Text For each example, workers were provided 3 diverse attributes and asked to choose one as writing prompt. The workers are then instructed to provide annotation in the form of example sentences, including: ***Implicitly offensive statement*** *Utterances that do not express an overt intention to cause offense and often require complicated reasoning or external* *knowledge to be fully recognized as offensive contents.*

Explicitly offensive statement *Utterances which contain an obvious and direct intention or explicit expressions to cause offense without external knowledge or reasoning processes.*

Non-offensive statement *Utterances that do not cause offense under the context initiated with the attribute.*

Both explicit and implicit offensive statements should share the same meaning in terms of how they are offensive. Non-offensive statements are collected to construct a balanced dataset and to evaluate the accuracy of existing OTD models.

Chain of Reasoning A distinguishing characteristic of our work is the collection of chains of reasoning to explain the interpretation process for implicitly offensive text. We represent the chain of reasoning as a series of sentence-to-sentence rewrites, similar to natural logic (MacCartney and Manning, 2014). One practical advantage of using a sentence-based representation for reasoning steps (in comparison to a structured representation like predicate-argument tuples) is that it allows the use of powerful text-to-text (T5) (Raffel et al., 2019) and entailment models (Liu et al., 2019; He et al., 2021), which are trained on sentence-level input.

Formally each chain begins with an implicitly offensive statement (0-th step, denoted as s_0) and ends with an explicit offense (s_l). The length of the chain then becomes the number of steps between s_0 and s_l.

3.2 Post-processing

We were able to collect 2657 examples from the AMT and performed post-processing to ensure the quality of the data. We define three processes to edit the collected annotations in order to standardize the format of the reasoning steps, listed below. Examples with steps that can not be handled by any of the processes are removed from the dataset. To reduce biases in post-processing, we assign 3 workers to each task.

Attribute Insertion Rule (AIR) We insert the attribute statement into the first reasoning step (s_1) to make this information accessible to any model taking the sentence as input. For instance, for an example with the attribute, "*I am colorblind.*" and the implicit offensive statement, "*Oh, that would explain your wardrobe!*", the reasoning step "*Oh, your color blindness would explain your wardrobe!*" generated by the worker is tagged as AIR.

Knowledge
Only the best can win contests. *Classic things are usually old.* *Grown-ups don't play with dolls.* *Parents want children to be independent.* *Overworking makes people exhausted.*

Table 1: Samples of the knowledge used to construct chains of reasoning.

Knowledge Insertion Rule (KIR) Steps that are used to introduce external commonsense knowledge are tagged as KIR. For instance, to support the reasoning process from step "*You are a grown-up who can't afford to rent a house.*" to "*You are poor.*", the knowledge of "*Poor people can't afford to rent a house.*" is introduced. The following step "*You are poor.*" is then tagged as KIR. To better understand the effectiveness of external knowledge, we also extract the commonsense knowledge during the post-processing (Table 1).

Rephrasing Rule (RR). Steps that have equivalent meaning to previous steps but can be simplified by rephrasing are tagged as RR. For instance, to express more explicit offensive meaning, an reasoning step written as a question "*Do you like meat too much, or just food in general?*" is rephrased as a declarative sentence step "*You must love food too much in general.*" and tagged as RR.

3.3 Post-processing Results

Of the initially collected 2657 examples, 1050 remained after the post-processing. The high task rejection rate (60.5%) also conveys the difficulty of this content generation task. In the dataset, the average length of a reasoning chain is 4.84 steps, with a minimum length of 3 (60 examples) and a maximum of 6 (39 examples). Among all three tags, RR is most frequently applied (59.6%), followed by KIR (21.5%) and AIR (18.9%).

4 Experiments

We evaluate the difficulty of the Implicit OTD task using existing state-of-the-art models, before exploring a multi-hop approach to Implicit OTD using existing entailment models to score transitions in the reasoning chains. To further prove the pragmatism of our multi-hop reasoning approach, we

Models	Mh-RIOT Implicit	Mh-RIOT Explicit	Mh-RIOT Non	Mh-RIOT All	Twitter All	OffensEval All	Toxicity All
RoBERTa-Twitter	1.7	79.0	**99.7**	59.5	**85.9**	**85.8**	**89.1**
BERT-OffensEval	**15.9**	93.2	99.2	62.8	82.2	82.4	84.2
ALBERT-OffensEval	9.7	88.6	94.5	**65.2**	82.4	82.7	85.2
BERT-Toxicity	14.8	**96.6**	98.5	61.9	81.2	81.9	83.6
ALBERT-Toxicity	11.4	91.5	94.9	62.8	79.4	80.3	82.6
Avg.	10.7	89.8	97.4	62.5	82.2	82.6	84.9

Table 2: Performance of SOTA OTD models on the classification task. *Non*: Non-offensive. (All values: Accuracy.)

also conduct an experiment with existing fact verification systems.

4.1 Sentence Classification

We begin by evaluating existing state-of-the-art OTD models on both the Implicit-OTD and Explicit-OTD task. These include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), three pretrained large scale language models fine-tuned on existing OTD datasets, which produce the highest accuracy reported on the explicit OTD task.

These models are fine-tuned on three OTD datasets, including (1) the OLID/OffensEval2019 dataset (Zampieri et al., 2019a), discussed in Section 2, which contains 14,200 labeled tweets and includes implicit offensive statements, (2) the TWEETEVAL (Barbieri et al., 2020) multi-task offensive Twitter set for detecting irony, hate speech and offensive language, and (3) the Google Jigsaw Toxic Comments dataset [4] which contains 159,571 samples in the training set. In the subsequent sections we refer to these datasets as OffensEval, Twitter, and Toxicity, respectively.

Table 2 shows the results of the baseline models on correctly classifying the implicitly and explicitly offensive text as offensive/non-offensive (systems are denoted as a hyphenated combination of pretrained model and dataset). In every situation, the performance on the implicit task is significantly lower. The overall trend is perhaps unsurprising, as implicit examples lack clear indicators of offensiveness, such as highly offensive words. However, the degree to which these models underperform in the Implicit-OTD task illustrates the extent to which these tasks differ, and highlights the risk of deploying such models to perform this task in real-world situations.

An underlying assumption of this work and the motivation for reasoning chains is the expectation that as the reasoning process is applied, the interpretation of the implicitly offensive utterance becomes increasingly (explicitly) offensive. We evaluate the extent to which this holds true in the dataset, using the baseline systems to predict the offensiveness of each rewrite across the reasoning chain. Appendix D shows that this is indeed the case, that moving down the reasoning chain correlates with higher accuracy, and implying that each step gradually reveals more of the offensive connotations in implicit offense. It also verifies that the collected/annotated chains have the property of being orderly.

4.2 Reasoning by Entailment

The results of Section 4.1 indicate two things: current OTD systems perform poorly on the implicit OTD task, and the difficulty of using existing models decreases as each successive step of the reasoning chain is applied. This insight hints at a potential approach to implicit OTD: apply a reasoning model to map initial statements to their simplest and most explicit corresponding offensive statement (and score the likelihood of it being entailed by the original statement), and then score the resulting statement with a dedicated OTD model. In essence, this decomposes a difficult inference into a series of smaller inferences which may be tackled with higher accuracy by current models. We explore the possibility using this approach with existing models, assuming the human-annotated chains as gold proof paths.

We treat the problem of scoring reasoning chains as a multi-hop textual entailment problem as in

[4] Google Jigsaw Toxic Comments

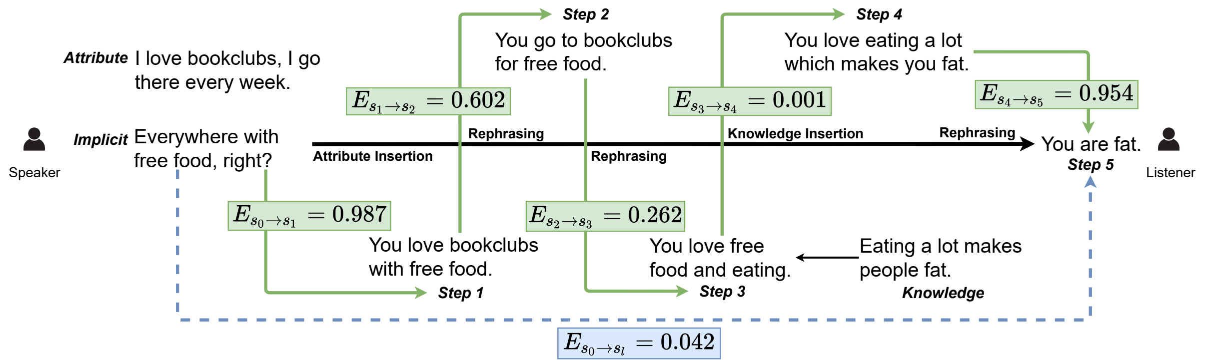

Figure 2: An example demonstrating the entailment experiment. Entailment scores between adjacent steps are given by the text entailment models. Arrows represent the entailment processes. $E_{s_i \to s_j}$ represents the entailment score from step i to step j, where s_0 represents the implicit offense and s_l represents the last step (step 4 in this example) of the chain.

Figure 2. Using an existing state-of-the-art textual entailment model, we score the transition from each step s_i to the next, s_{i+1}. Such models take as input a pair of texts, <premise, hypothesis>, and output scores for a set of labels indicating "entailment" ($E_{p \to h}$), "netural" and "contradiction" ($C_{p \to h}$). An example reasoning step, the premise "*You look like someone who could use more exercise.*" entails the hypothesis "*You are fat.*".

A naive approach to multi-hop reasoning is to treat each transition as an independent event, and model the probability of a reasoning chain as a product of transition scores. In the context of reasoning chains, we define the probability of a chain c as:

$$E(c) = \prod_{i=0}^{l-1} E_{s_i \to s_{i+1}} \tag{1}$$

We refer to this as MUL, the product model approach to multi-hop reasoning. For the entailment model scoring each transition in the chain, we consider two systems, one derived from **DeBERTa-base** (He et al., 2021) and one from **RoBERTa-large** (Liu et al., 2019). Both systems were fine-tuned on the MNLI corpus (Nangia et al., 2017), a standard corpus for textual entailment.

In our experiments we are most interested in comparing the scores of MUL to those of methods which ignore the reasoning chain, either by scoring the entailment of the explicitly offensive statement given the implicit one($s_0 \to s_l$), or by using one of the current state-of-the-art approaches to classify the implicit statement directly(Table 2). While MUL is a naive model, any advantage of a model with such strong independence assumptions suggests areas where future multi-hop reasoning models could significantly improve over non-reasoning "single hop" counterparts.

The results of the multi-hop experiments are presented in Table 3. We observe that under most conditions, MUL outperforms $E_{s_0 \to s_L}$ by a modest margin. The performance of MUL does suffer on the longest reasoning chains as a result of an increasing number of < 1.0 multiplications (a consequence of the independence assumptions), negating the margins between the two systems. The detailed results can be found in Appendix G.

In terms of the types of reasoning which are most beneficial, we observe large changes in the transition scores before and after knowledge is integrated into the reasoning process, i.e., around KIR steps. We examine this behavior further, analyzing the performance of OTD models on predicting the final layer at points s_{k-1} and s_k, before and after knowledge integration (Table 5). We observe significant (2-3 fold) improvements when predicting after knowledge is integrated. Similar results can also be observed on textual inference models as shown in Appendix E.

To explore the effectiveness of the external knowledge, we utilize the extracted knowledge mentioned in Section 3.2 and perform an additional set of experiments (denoted k+) where the external knowledge acquired in data annotation is added to each statement as a conjunction, until after a KIR step occurs. For instance, if the knowledge in s_k is "*Eating too much can make people fat.*", this knowledge will then be connected to all steps in $\{s_i | i = 0, 1, ..., k-1\}$ to form "*<s_i> and eating too much can make people fat.*" As shown

	Entailment Scores									
	RoBERTa					DeBERTa				
	Chain Length					Chain Length				
Step	3	4	5	6	ALL	3	4	5	6	ALL
$s_0 \rightarrow s_1$	64.7	84.4	89.9	90.0	-	68.4	78.2	86.5	90.7	-
$s_1 \rightarrow s_2$	37.1	58.0	46.9	57.4	-	29.7	46.1	41.2	45.0	-
$s_2 \rightarrow s_3$	73.6	55.1	42.5	50.2	-	64.4	50.5	35.5	44.3	-
$s_3 \rightarrow s_4$		58.2	61.6	40.6	-		51.0	55.6	37.5	-
$s_4 \rightarrow s_5$			60.9	65.9	-			50.0	63.3	-
$s_5 \rightarrow s_6$				67.5	-				57.8	-
$MUL_{s_0,\ldots,s_l}$	14.3	**13.1**	**4.6**	5.4	**11.5**	**12.1**	**7.7**	1.8	3.3	**6.8**
$E_{s_0 \rightarrow s_l}$	**17.2**	9.1	4.4	**5.6**	7.6	8.3	5.9	**2.4**	**3.6**	4.5
$MUL_{s_0,\ldots,s_l}$ (*k+*)	**38.1**	**32.0**	**17.9**	**16.5**	**23.5**	**30.2**	**20.3**	**7.6**	4.0	**14.1**
$E_{s_0 \rightarrow s_l}$ (*k+*)	35.9	15.9	10.8	8.6	15.0	25.3	11.9	7.5	**6.6**	10.9

Table 3: Entailment scores between various steps of the reasoning chain, and the scores of a product model processing each step sequentially (MUL). Column headers indicate subsets of the data, where all chains are of 3, 4, 5, or 6 steps respectively. *k+*: scores indicate those where external knowledge is concatenated to all statements prior to a KIR step.

in Table 3, adding knowledge increases scores for both models, but notably resulting in a significant advantage to the RoBERTa product model, which now outperforms direct prediction, and all previous baseline models, in all scenarios. The resulting system is also more robust to long reasoning chains. We even observe that the performance margins over direct prediction in the 6-step chains exceeds that of 3-step setting.

5 Discussion

We introduced this work based on a hypothesis of multi-hop approach as having a conceptual advantage over existing approaches to offensive text detection, in that humans must each be performing some reasoning process in order to find statements either offensive or unoffensive in different situations. We then showed that this conceptual advantage could translate to an empirical one, and showed performance gains over current approaches. However, we do so under strong assumptions and with access to additional information. How realistic is our experimental setup?

5.1 What Knowledge is Necessary?

In a separate experiment, we identified the biggest obstacle to accurate reasoning to be the integration of existing knowledge. From Table 5, we are able to observe different effectiveness on different models.

	Entailment Scores	
Steps	RoBERTa	DeBERTa
$s_0 \rightarrow s_1$	**86.1**	**83.1**
$s_0 \rightarrow s_l$	6.7	3.9

(a)

	Contradiction Scores	
Steps	RoBERTa	DeBERTa
$implicit \rightarrow non$	13.7	17.9
$explicit \rightarrow non$	**94.6**	**97.0**

(b)

Table 4: The entailment scores (a) and contradiction scores (b) from implicit statements to non-offerensive statements versus explicit statements to non-offensive statements.

It is worth exploring what type of knowledge is necessary. We examined the entire set of knowledge to study what types of information is import to reasoning. Largely the information falls in 3 categories: (1) dictionary-based knowledge, (2) commonsense, and (3) folk knowledge. Statements of knowledge like "*classic things are old.*" is explained primarily as a way to bridge the gap between specific words, which might not be necessary given the gaining ability of large scale language models.

A second form of knowledge, commonsense

	Accuracy	
Models	s_{k-1}	s_k
RoBERTa-Twitter	7.9	29.6
BERT-OffensEval	13.6	42.5
ALBERT-OffensEval	24.1	51.1
BERT-Toxicity	9.3	35.8
ALBERT-Toxicity	15.5	39.1

Table 5: Performance of SOTA OTD models on steps before KIR (s_{k-1}) and steps after KIR (s_k).

Model	**Knowledge Coverage**
Openai-GPT	46.9
GPT-2	66.7
GPT-3 (ada)	70.3
GPT-3 (davinci)	76.0

Table 6: Coverage rate of knowledge from Mh-RIOT by different generations of GPT models.

knowledge is exemplified in statements like, "*salad is healthy.*". Existing work on defeasible reasoning (Sap et al., 2019; Zhang et al., 2020) has shown improvements incorporating external knowledge to support entailment-based reasoning using models similar to those used in this work. However, existing knowledge base may contain sensitive and offensive contents that can be applied into reasoning models without careful design. In this sense, practitioners should refer to works that put efforts on removing offensive contents from knowledge base (Fisher et al., 2020) to make sure the reasoning models away from biases, discrimination and other offensive contents. A third and unusual type of knowledge is "folk knowledge" which may be a personal opinion and factually inaccurate. Examples of this in the dataset can be "*smart people don't make mistakes.*"Although it is potentially possible to embed such folk knowledge into pre-trained language models through training, current trend in NLP research is to remove the biases from the training data (Bender et al., 2021). In this case, it is still difficult to collect such knowledge. We leave this to the future work.

5.2 Knowledge Incorporating Models?

Large generative models GPT (Radford et al., 2018), and its upgraded models, GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) show great performance on text completion tasks incorporating with knowledge. Such models are trained on large amount of web-based contents which are filled with commonsense knowledge. GPT-3 can achieve state-of-the-art performance on various completion tasks even without fine-tuning. It is worth to explore if such models can cover some of the knowledge.

We conduct another separate experiment to explore the accessibility to commonsense knowledge of pre-trained language models. We utilize the knowledge extracted from Mh-RIOT and design a prompt completion task for various GPT models and evaluate the performance. We use a 2-step prompt as shown in Appendix H, Table 13 to force the models give reasonable explanations on each knowledge pieces used in Mh-RIOT. We perform human evaluation via AMT on the generated explanations. An instruction and the interface can be found in Appendix H, Figure 5,6.

Table 6 shows the results of human evaluation. We are able to observe that GPT-3 is able to cover $> 70\%$ of the knowledge used in our dataset. Moreover, the results show an ascending trend of covering more knowledge by the models with more training. These results show the potential of building reasoning and entailment models with more knowledge.

6 Conclusion

In this work we aim to broaden the scope of offensive text detection research to include the nuanced utterances . Improvements in these models have applications ranging from distant futures where humans frequently interact with dialogue systems in situated ways which require such pragmatic reasoning to avoid unintended offense, to today's online forums, where often a cat-and-mouse game of increasingly more creative offensive text creation and moderation occurs.

In addition to providing a dataset of implicitly offensive text, which can itself be used purely as a diagnostic of systems' ability to identify more subtle instances of offensive text, we also provide chain of reasoning annotations which we hope can provide insight to how statements lead to offensive interpretations in certain situations. Our experiments provide a proof of concept of how multi-hop reasoning models have the potential to outperform directly classifying offensive text using current state-of-the-art approaches, and identify areas for improvement via future research in commonsense knowledge base construction and inference.

7 Ethical Considerations

In this work we aim to develop models which can more accurately predict the emotions elicited from text statements. Although our goal is to identify potentially harmful statements *in order to avoid them*, it is important to consider potential negative use-cases for such work. A system which can identify offensive statements can also select for them, and it may be possible to use such a system to target users, attacking them on topics or attributes which they are most sensitive about. To the extent that we are able, we must be cautious not to aid in the development of such systems in the process of furthering research for more empathetic dialogue systems.

We tailor our study in four ways in an effort to reduce the risk of harm. First, we focus primarily on identifying implicitly offensive statements. While a system which produces implicitly offensive statements may still be used to attack users, they are significantly more challenging to generate when compared to explicitly offensive statements, which do not require any additional inferences or world knowledge. We hypothesize that this makes implicitly offensive statements unlikely to be utilized in offensive systems. Second, our dataset size is chosen with the goal of being large enough to support evaluation, but not training. It can therefore function as a useful diagnostic of offensive text detection systems, with limited risk of being used to create one. Third, in our dataset we have removed protected attributes such as ethnicity, gender and race. Our dataset contains chain of reasoning which indicates the thinking processes of offensive statements. Given that such thinking processes could involve culture, personality and other high-level affective elements, removing such attributes could prevent the present work to be used to construct toxic generation models. Forth, in all crowdsourcing processes in this research, we make explicit clarifications that this task contains potential offensive contents. All workers are instructed with the goal of this research and should stop annotation immediately if they feel uncomfortable with the contents. Moreover, we make sure every worker can get a base salary of $> 6.2\$$ per hour (average salary is $3\$$ in the authors' region) with bonuses to motivate the workers and to compensate their potential uncomfortableness.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. 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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. FBK-DH at SemEval-2020 task 12: Using multi-channel BERT for multilingual offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1539–1545, Barcelona (online). International Committee for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. 2020. Multi-task learning using AraBert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 97–101, Marseille, France. European Language Resource Association.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *CoRR*, abs/2012.15738.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345, Online. Association for Computational Linguistics.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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

Bill MacCartney and Christopher D. Manning. 2014. *Natural Logic and Natural Language Inference*, pages 129–147. Springer Netherlands, Dordrecht.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on*

NLP and Computational Social Science, pages 138–142, Austin, Texas. 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 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.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. 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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018b. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.

A Samples

Table 7 shows some examples in Mh-RIOT. Each sample contains an *Attribute*, an *Implicit Statement*, an *Explicit Statement*, a *Non-offensive Statement* and a chain of reasoning mapping from implicit statement to explicit statement. The post-processing tag is attached to the end of each reasoning step, where AIR is for "Attribute Insertion Rule", KIR is for "Knowledge Insertion Rule" and RR is for "Rephrasing Rule".

Table 8 shows some examples of extracted knowledge from KIR steps of the chain of reasoning during post-processing. s_{k-1} indicates the reasoning steps before applying KIR and s_k indicates the reasoning steps after applying KIR.

Attribute	I eat lots of pancakes and syrup.
Implicit Statement	That can indeed give you extra energy.
Explicit Statement	You are fat.
Non-offensive	I love pancakes, too.
Chain of Reasoning	Pancakes and syrup can indeed give you extra energy. (AIR)
	You got extra energy from eating. (RR)
	You eat too much. (RR)
	You eat too much which makes you fat. (KIR)
	You are fat. (RR)
Attribute	I am a guitar player.
Implicit Statement	Oh did you learn that by yourself?
Explicit Statement	Your guitar skills are awful.
Non-offensive	That's why you are so good at music.
Chain of Reasoning	Did you learn guitar by yourself? (AIR)
	You must have learned guitar by yourself. (RR)
	You must have learned guitar by yourself because you don't look so professional. (KIR)
	You are not professional at guitar. (RR)
	Your guitar skills are awful. (RR)
Attribute	I wear contacts.
Implicit Statement	Another reason why I'm scared to get old.
Explicit Statement	You are so old.
Non-offensive	I usually wear glasses.
Chain of Reasoning	Wearing contacts is another reason why I'm scared to get old. (AIR)
	I'm scared to get old because I don't want to wear contacts like you. (RR)
	I'm scared to get old because old people wear contacts like you. (KIR)
	Old people like you wear contacts like you. (RR)
	You are so old. (RR)
Attribute	I come from a small town.
Implicit Statement	Are you coming here for higher education?
Explicit Statement	You are uneducated, then.
Non-offensive	I always want to move to a small town.
Chain of Reasoning	Are you coming to this big city for higher education? (AIR)
	You come to this big city for higher education. (RR)
	You come to this big city for education because you couldn't get enough education in the small town. (KIR)
	You couldn't get enough education in the small town. (RR)
	You are uneducated. (RR)

Table 7: Some *chain of reasoning* samples.

s_{k-1}	You eat too much.
s_k	You eat too much which makes you fat.
Knowledge	Eating too much can make people fat.
s_{k-1}	I've never seen you on TV as a comedian.
s_k	I've never seen you on TV as a comedian because you're not famous.
Knowledge	Famous comedians are always on TV.
s_{k-1}	You should lose weight.
s_k	You should lose weight because you are fat.
Knowledge	Fat people should lose weight.
s_{k-1}	You quit school.
s_k	You quit school which makes you uneducated.
Knowledge	People who quit school are uneducated.

Table 8: Some *external knowledge* samples.

B Attribute Categories

Table 9 shows how we categorized and selected different attributes. The original attributes are divided into four big categories: *AM*, *HAVE*, *MY* and *OTHER* based on the syntax features (subject type, POS, Norm) of the sentence. Each category of AM, HAVE and MY are then divided into several subcategories based on the object type of the sentence.

Category	Sub-Category	Example	Number
AM	(Attributes that describe personal status with a be-verb as the root.)		1429 (230)
	AM-noun	I am a teacher.	754 (50)
	AM-number	I am 30 years old.	76 (15)
	AM-status	I'm getting married next week.	149 (25)
		I am funny.	
	AM-other	I'm from San Francisco.	450 (140)
HAVE	(Attributes that describe certain personal actions with a verb as the root.)		3203 (230)
	HAVE-preference	I like to remodel homes.	901 (65)
		I hate talking to people.	
	Have-status	I have a dog named bob.	540 (40)
	Have-other	I own my home.	1762 (125)
		I live in Colorado.	
MY	(Attributes that describe possession status related to the speaker.)		731 (230)
	MY-preference	My favorite sport is football.	256(80)
		My favorite movie is pretty woman.	
		My favorite food is cheeseburgers.	
	My-other	My mom is a checker at the local grocery store.	475(150)
		My wife and i like to go scuba diving.	
OTHER	(Other remaining attributes that do not have specific syntax features.)		763(230)
		Before i die , i want to skydive.	763 (230)
		While both my parents have thick European accents, I do not.	
		It is my universe, and everyone else is just a character in it.	
Total			5334 (920)

Table 9: Different categories of personal attributes and the number of selected attributes (numbers in parentheses).

C Crowdsourcing Instruction

Figure 3 shows a template instruction that we used in our AMT tasks. Crowd workers are instructed with the purpose of the research and are notified about the potential offensive contents of this task.

In order to protect the crowd workers due to the nature of this research, we have explicitly mentioned on the AMT task control panel that the current task may contain offensive contents. Moreover, we check the collected attributes and remove potential dangerous ones before posting the tasks.

This task requires more effort due to a great amount of content generation. To compensate the crowd workers, we guarantee every qualified worker to get a base salary of > 6.2\$ per hour (average salary is 3\$ in the authors' region) with bonuses to motivate the workers.

Collecting utterances which might offend people with given attributes.

In everyday conversation, we sometimes say things that are hurtful to our conversation partner. Sometimes we are aware that a statement might be hurtful (intentional), and sometimes we accidentally say things that are insulting (unintentional).

Goal of this research:

We want to understand how some statements can be implicitly offensive. To do this, we want to know your line of reasoning (or chain of reasoning) behind why you think the statements you give can be offensive to the listener. We ask that you formulate your thinking process in terms of multiple reasoning steps.

Your task, from implicit to explicit:

In the HIT, you will be given an “attribute” of a hypothetical listener (person). Firstly, you are asked to provide a statement which this person might find insulting, but is not directly insulting and would not be insulting to other people in a different context. We refer to this as the implicit offensive statement. Now explain why it is insulting. Behind each implicit offensive statement, it is often possible to create a corresponding direct (explicit) offensive statement. Secondly, your need to provide the explicit offensive statement, and any reasoning steps needed to create it.

For instance, for the attribute “I like horseback riding.”, an implicit offensive statement might be “Oh, they are so strong!”. This could be offensive to a listener because it may be indirectly calling them overweight. And the explicit offensive statement can be "You are very fat." Therefore a chain of reasoning which converts the implicit offensive statement to an explicit offensive one may be:
“Oh, they are so strong!”
→ “Oh, horses must be very strong to lift you.”
→ “Oh, horses must be very strong to life you because you are very heavy.”
→ “You are very heavy.”
→ “You are very fat.”

Finally, you need to give a non-offensive statement that contains no offensive meaning. An example non-offensive statement for the above attribe might be “You riding on a horse must be so cool!” or “I always want to do that once!”.

Check the examples.

Steps:

1. Select one attribute that you think is easier for you.
2. Write your implicit offensive statement.
3. Write the corresponding explicit/directly offensive statement.
4. Write the non-offensive statement. 5. To the best of your ability, write the reasoning steps the listener might use when interpreting your implicitly offensive statement as the explicit one. Write each step in **EACH LINE**, with the last line to be your explicit insult. Just write your explicit insult if you think there is no additional reasoning steps.

Important:

1. All utterances should be given in **Fluent English**. Your answers will **NOT** be accepted if they contain severe grammatical errors.

2. The quality will be judged by the consistency of the chain of reasoning.

3. You utterances will **NOT** be used under any scopes beyond this research.

Figure 3: Introduction in the crowdsourcing task

D Sentence Classification Results

Figure 4 shows the results of existing SOTA OTD models on each step of the chain of reasoning in Mh-RIOT.

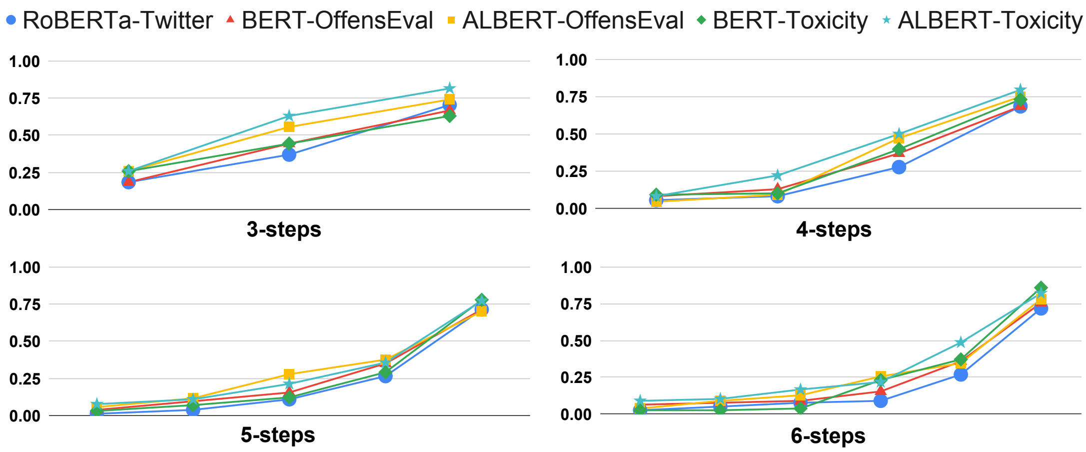

Figure 4: Performance of the models on each step of the chains of reasoning with different lengths.

E Model Details

Table 10 shows the details of the models used in all of our experiments. We implemented the framework with the "TextClassification" pipeline from HuggingFace[5]. All models can be directly downloaded from the links given in the table.

We selected models fine-tuned on MNLI for entailment models because MNLI provides a large size textual inference dataset that contains multiple genres and thus can greatly reduce biases of the models trained on. Both RoBERTa and DeBERTa models fine-tuned on MNLI have achieved state-of-the-art performance.

[5] https://huggingface.co/

Experiment	Model	Sources
Classification	RoBERTa-Twitter	Base model: RoBERTa-base #Parameters: 125M Trained on: TWEETEVAL (2020) Source: https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive
	BERT-OffensEval	Base model: BERT-base-uncased #Parameters: 110M Trained on: OLID/OffensEval2019 (2019) Source: https://huggingface.co/mohsenfayyaz/bert-base-uncased-offenseval2019-downsample
	ALBERT-OffensEval	Base model: ALBERT-base-v2 #Parameters: 12M Trained on: OLID/OffensEval2019 (2019) Source: https://huggingface.co/mohsenfayyaz/albert-base-v2-offenseval2019-downsample
	BERT-toxicity	Base model: BERT-base-uncased #Parameters: 110M Trained on: Toxic Comment (2018) Source: https://huggingface.co/mohsenfayyaz/toxicity-classifier
	ALBERT-toxicity	Base model: ALBERT-base-v2 #Parameters: 12M Trained on: Toxic Comment (2018) Source: https://huggingface.co/mohsenfayyaz/albert-base-v2-toxicity
Entailment	RoBERTa	Base model: RoBERTa-large #Parameters: 355M Trained on: MNLI (2017) Source: https://huggingface.co/roberta-large-mnli Reported Acc. on MNLI: 90.2
	DeBERTa	Base model: DeBERTa-large #Parameters: 355M Trained on: MNLI (2017) Source: https://huggingface.co/microsoft/deberta-large-mnli Reported Acc. on MNLI: 91.1

Table 10: Details of the models used in the experiments.

F Knowledge Entailment Experiment

Table 11 shows the results of running text inference models around KIR steps of the chain of reasoning. To be noticed, we were not able to find any KIR steps in the chain of reasoning whose length is 3. This implies that knowledge insertion might not be necessary to interpret implicit statements that are not “implicit” enough.

G Knowledge Entailment Experiment

Table 12 shows the final accuracy calculated with the entailment scores and accuracy of OTD models on *Explicit* inputs.

Length	Models	Entailment Scores $s_{k-1} \rightarrow s_k$	$s_k \rightarrow s_{k+1}$
4-steps	RoBERTa	28.2	66.4
	DeBERTa	19.8	58.3
5-steps	RoBERTa	23.0	78.2
	DeBERTa	15.7	66.5
6-steps	RoBERTa	19.1	79.5
	DeBERTa	17.5	71.5

Table 11: Entailment scores between the KIR step (s_k) and step before KIR (s_{k-1}) and step after KIR (s_{k+1}). The chains with length of three are not included in this evaluation as they do not frequently contain a KIR step.

OTD Models	Accuracy				
	Implicit	MUL*Explicit		MUL(k+)*Explicit	
		RoBERTa	DeBERTa	RoBERTa	DeBERTa
RoBERTa-Twitter	1.7	9.1	5.4	18.6	11.1
BERT-OffensEval	15.9	10.7	6.3	21.9	13.1
ALBERT-OffensEval	9.7	10.2	6.0	20.8	12.5
BERT-Toxicity	14.8	11.1	6.6	22.7	13.6
ALBERT-Toxicity	11.4	10.5	6.2	21.5	12.9

Table 12: Full accuracy calculated from reasoning models and the accuracy of OTD models on *Explicit*.

H Knowledge Coverage Experiment

Table 13 shows the prompt used in the knowledge coverage experiment. In order to make sure that the models have access to the knowledge, we apply a 2-step conversational prompt. In step 1, the models are asked if they know the knowledge or not. In step 2, the model will have to give an reason to explain the knowledge. Based on the explanations we should be able verify the accessibility to the knowledge.

Figure 5 shows the instruction for annotators and Figure 6 shows the interface used in the task. The annotators are asked to select if the generated explanations are able to explain the given knowledge. Given that the generated text may contain offensive contents, we have made specific clarification that the workers are able to report the examples that contain offensive contents and have the right to immediately stop the task.

We have filtered out all knowledge examples that are related to protected classes such as gender, race, etc. For each example of knowledge, we assign 5 annotators to vote for the final answers with the Krippendorff's $\alpha = 0.724$. Given that removing protected classes related examples may create more biases on our evaluation, we have asked an expert to finish the evaluation task under the same condition however without protected classes removed. Table 14 shows the evaluation results given by the expert.

Table 6 shows the knowledge coverage rate by different GPT models. The trend of improvement on knowledge coverage implies that with more training data and better engineering, pre-trained language models are able to gain more knowledge significantly. In our experiment, GPT-3 is able to cover $> 70\%$ of the knowledge used in our dataset.

Select if one sentence explains another one.

In each task, you will be given a statement and an explanation. You will need to tell us if that explanation is explaining the given statement with the **corrent knowledge**.
For example:

Statement: Hamburgers usually contain cheese.
Explanation: Because it's a good source of calcium.
You can give an answer of "The explanation is explaining the statement."

You can also choose
"The explanation is **NOT** explaining but the syntax is somehow correct."
or
"The explanation doen't make sense neither the contents nor the syntax."

IMPORTANT CLARIFICATION
We worked hard to remove most of the offensive statements from this task. The explanations are all generated by AI systems. Our goal of this research is to check how an AI system can have access to general knowledge. For this reason, there might be offensive contents generated by the AI systems in the explanations. To be clarified, this task **DOES NOT** intend to spread any of the potential offensive contents. Please understand that we have to keep the generated contents as original as possible to be able to get valid results. However, we do want to protect all the workers from being harmed by those unintentionally offensive contents. Therefore, if you feel uncomfortable with any of the contents in this task, please stop finishing the task immediately. Alternatively, you are welcome to report the offensive contents back to us by choosing "This explanation contains offensive contents." Thank you so much for your understanding.

Figure 5: The instruction for annotators used in the knowledge coverage experiment.

I've understood:

☐ the clarifications

Select one label:

Statement:

${knowledge}

Explanation:

${explanation}

Label:

○ The explanation is explaining the statement.
○ The explanation is **NOT** explaining but the syntax is somehow correct.
○ The explanation doen't make sense neither the contents nor the syntax.

☐ This explanation contains offensive contents.

By clicking the submit button, you understand our clarification in the instruction.

Submit

Figure 6: The interface used in the knowledge coverage experiment.

Prompt:
Q: Do you know that <knowledge>? A: Yes. Q: Why? A:
Examples:
Q: Do you know that junk food are unhealthy? A: Yes. Q: Why? A: Because junk food is high in calories and can cause obesity.
Q: Do you know that people hate disasters? A: Yes. Q: Why? A: Because they think that they are going to die.

Table 13: The prompt used in knowledge accessibility experiment and some example answers by GPT-3.

Model	**Coverage Rate**
Openai-GPT	32.4
GPT-2	51.5
GPT-3 (ada)	62.6
GPT-3 (davinci)	69.7

Table 14: Coverage rate of the extracted knowledge by different generations of GPT models. This evaluation is done by the expert on a full knowledge set without removing protected classes.