Rethinking Offensive Text Detection as a Multi-Hop Reasoning Problem

Anonymous ACL submission

Abstract

We introduce the task of implicit offensive language detection in dialogues, where a statement may have either an offensive or unoffensive interpretation, depending on the listener and context. We argue that inference is crucial for understanding this broader set of offensive utterances, and create a dataset featuring chains of reasoning to describe how an offensive interpretation may be reached. Experiments show that state-of-the-art methods of offense classification perform poorly on this task, achieving less than 0.12 average accuracy. We explore the use of pre-trained entailment models as part of a multi-hop approach to the problem, showing improved accuracy in most situations. We discuss the feasibility of our approach and the types of external knowledge necessary to support it.

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, opinions, or points-of-view. However, their benefit to society is often marred by 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. Work in NLP has sought to automate the identification of toxic speech, and has achieved high accuracy in specific domains, such as identifying sexist (Golbeck et al., 2017), racist (Waseem, 2016), or otherwise hateful text (Ross et al., 2016; Gao and Huang, 2017; Davidson et al., 2017).

While many instances of toxic speech on the web 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 abusive text**, which is unambiguous in its potential to be offensive and often includes overtly offensive terms, such as slurs, and (2) **implicit abusive text**, which is more ambiguous, and may use sarcasm, innuendo, or other rhetorical devices to hide the intended nature of the statement. Previous ML-based approaches to offensive text detection deal almost exclusively with explicit text detection, and achieve high accuracy on many domains. The large pre-trained language models utilized in state-of-the-art offensive text detection systems can exhibit a remarkable ability to infer and reason about the true meaning of text, and so in this work we ask: *how effective are these approaches when applied to implicit offensive text detection? Are other methods required to perform this task well?*

We begin by formalizing the task of implicit offensive text detection. Waseem et al. define implicit abusive text, but they do not discuss the relationship between implicit and explicit offensive text. In this work we argue that each implicitly offensive statement is offensive because it has a corresponding explicitly offensive statement, which is closer to the sentiment the listener feels when interpreting the statement as offensive. Consider the dialogue between two speakers, S1 and S2:

S1: “I love bookclubs, I go every week”
S2: “Do they have free food or something?”

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

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

interpretation to be a crucial part of recognizing implicitly offensive statements. As an extreme case of this, consider statements which are unintentionally offensive, or that the same statement may be considered either offensive and unoffensive depending on who interprets it (and in what context).

To study this phenomenon 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). We evaluate state-of-the-art offensive text detection models on explicit offensive text and reaffirm that they are able to perform the task with high accuracy, sometimes achieving $> 90\%$. However, when used for implicit detection, their accuracy drops to an average of 10%. We then explore the use of reasoning-based approaches to the solution, using currently available textual entailment models to score each reasoning step in the chain. Even when using strong independence assumptions (treating each step as an independent event, and therefore scoring each chain as a product of reasoning step probabilities), the multi-hop reasoning approach performs comparable, and in some cases better than, state-of-the-art models. We examine the role that external knowledge plays in the reasoning process, and identify future directions for dedicated reasoning systems for offensive text detection.

Our contributions in this work are threefold:

- We propose the task of implicit offensive text detection, and collect a dataset to support research on this topic (with additional annotations for reasoning-based approaches).
- We conduct experiments using existing state-of-the-art offense detection models, and show they perform poorly when tasked with predicting implicit offensive text.
- We examine the use of existing entailment models as part of a multi-hop reasoning approach to implicit textual offense detection. We provide an analysis of where reasoning succeeds, where it fails, and what types of external resources would be necessary to support reasoning-based approaches for offensive text detection.

2 Related Works

Offense Detection in Text Classification Early approaches to offensive language detection rely primarily on dictionaries like hatebase [2] to filter offensive words and phrases. Early machine learning-based approaches utilized simple features, such as bag-of-word representations, to train models from small datasets (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 non-English languages (such as Italian, (Rizwan et al., 2020), Arabic (Mubarak et al., 2020; Chowdhury et al., 2020; Husain and Uzuner, 2021), Greek (Pitenis et al., 2020)). Supported by larger datasets, a number of deep learning-based methods have been proposed (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.

Offense Detection in Dialogue Offensive text detection in dialogue is an important problem since dialogue systems trained on toxic content may reproduce it in interactions with human users. This problem has previously been studied in the context of human-in-the-loop system improvements (the “Build it Break it Fix it” paradigm (Dinan et al., 2019)), which found that the offensiveness of the statement must be determined within the context of the larger dialogue (similar to the motivation of this work). Other dialogue-specific work on identification of offensive text includes detecting toxic comments (Gehman et al., 2020a), gender bias (Dinan et al., 2020) and racism (Zhou et al., 2021). Dialogue-based datasets for offensive text detection also exist (Cercas Curry and Rieser, 2018), though to our knowledge, we are the first to provide a dataset test for implicit offensive text detection with reasoning chains. Detoxifying language can also occur during generation (rather than during training or as a data cleaning step during pre-processing) (Krause et al., 2020; Gehman et al., 2020b), and our dataset could be used as an additional challenge dataset and diagnostic tool for these systems.

Reasoning Processes of Offense The Offensive Language Identification Dataset (OLID) is one of the most commonly used datasets for offensive

[2] www.hatebase.org

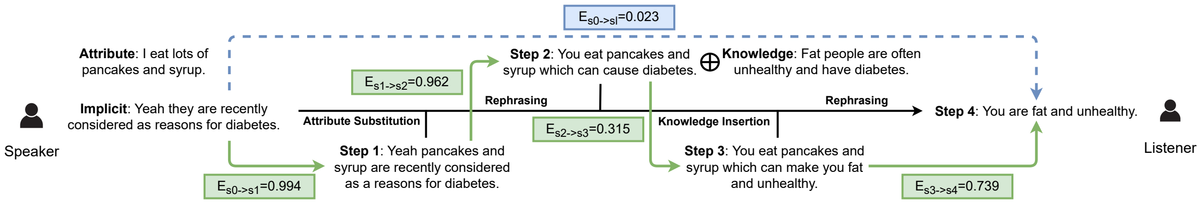
E_{s0->sl}=0.023
Attribute: I eat lots of pancakes and syrup.
Step 2: You eat pancakes and syrup which can cause diabetes.
⊕
Knowledge: Fat people are often unhealthy and have diabetes.
E_{s1->s2}=0.962
Implicit: Yeah they are recently considered as reasons for diabetes.
Speaker
Rephrasing
Attribute Substitution
E_{s2->s3}=0.315
Knowledge Insertion
Rephrasing
Step 4: You are fat and unhealthy.
Listener
Step 1: Yeah pancakes and syrup are recently considered as a reasons for diabetes.
E_{s0->s1}=0.994
Step 3: You eat pancakes and syrup which can make you fat and unhealthy.
E_{s3->s4}=0.739

Figure 1: An example demonstrating the entailment experiment. Entailment scores between adjacent steps are given by the text entailment models. Arrows represent the entailment processes. $E_{s_i \rightarrow s_j}$ 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.

text detection (Zampieri et al., 2019a,b, 2020), and consists of a multi-level annotation scheme. Each level dictates the target of the offensive text, in terms of their identity as a group, individual, or entity. Caselli et al. (2020) augmented the OLID with labels for capturing the degree of explicitness in the offense (defined primarily as the presence of an overtly offensive word/slur), but obtaining a significant number of truly ambiguous implicit offensive statements is difficult enough task that we provide our dataset as a dedicated resource for this task, therefore guaranteeing the presence of some chain of reasoning to a corresponding explicitly offensive statement. In this sense, a more similar approach comes from normative reasoning in moral stories (Emelin et al., 2020), where the focus is to predict the "moral norm" with a two-hop style input of "moral/immoral action" and "moral/immoral consequence".

3 Data Collection

The dataset consists of three parts: (1) a personal attribute the reader/listener has (thus providing some context in which to interpret the potentially offensive statement), (2) an implicitly offensive statement implicit and its corresponding explicitly offensive statement, and (3) chain of reasoning for describing the relationship between the two offensive statements.

Mechanical Turk was used to collect 2,800 examples, of which 1,000 remained after filtering for quality.

3.1 Personal Attribute

The goal of dataset construction is to create a collection of implicitly offensive statements for further study, and as we have defined in Section 1, the context in which a statement occurs is crucial to understanding its interpretation as offensive. As it can be difficult to ask annotators to provide statements which are ambiguously offensive and relevant to an existing dialogue, we reduce the context to a single feature: a personal attribute of the reader/listener. By introducing attributes, we are able to: 1) limit the domain of generated utterances, 2) establish context for further reasoning. We collect a set of attributes from the profile sentences in the PERSONCHAT corpus (Zhang et al., 2018a), of the form "*I like sweets.*", or "*I work as a stand up comedian.*". Attributes related to ethnicity, gender, sexuality, and other protected classes are manually removed, leaving 5334 distinct attributes. 350 attributes were chosen for use in the dataset, in order to have multiple annotations for each attribute.

3.2 Implicit & Explicit Text Pairs

For each given attribute, we collect two different types of offensive statements, the *implicitly* offensive statement and the corresponding *explicitly* offensive statement, as defined below:

Implicit 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.*

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

We ask each annotator to provide an implicitly offensive statement (which would be offensive to a reader who has the given attribute), after which they are asked to rewrite the utterance as an explicitly offensive statement so that the both statements share the same meaning in terms of being offensive.

3.3 Chain of Reasoning

A distinguishing characteristic of our work is the use of chains of reasoning to explain the interpretation process for implicitly offensive text. We repre-

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.

sent the chain of reasoning as a series of sentence-to-sentence rewrites. One practical advantage of choosing a sentence-based representation for the reasoning steps 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 is not immediately compatible with structured representations like predicate-argument tuples. Each chain begins with an implicitly offensive statement (0-th step, denoted as s_0) and ends with an explicit offense (s_l), making the length of the chain the number of steps between s_0 and s_l, inclusive.

3.4 Annotation Guidelines

The high annotation rejection rate (64.3%) conveys the difficulty of this particular annotation task. We utilize common tactics for improving annotation quality, including performing annotations in batches, and removing poor annotators from future data iterations. We employ a number of additional annotation guidelines to help normalize the collected annotations, applied in a second stage by a different set of annotators, after the first round had commenced.

Attribute Substitution Rule (ASR). This rule allows annotators to substitute part of the implicit offense with the given attribute. ASR is often used to create the 1-st step (denoted as s_1) of the chain which asserts that the chain of reasoning can be consistent with the context given in the attribute. For instance, regarding the attribute "*I am color-blind.*" and the implicit offense "*Oh, that would explain your wardrobe!*", the s_1 in the chain can be created with the ASR as "*Oh, your color blindness would explain your wardrobe!*"

Knowledge Insertion Rule (KIR). This rule allows annotators to insert commonsense knowledge to support the reasoning. Table 1 shows some samples of the external commonsense knowledge used by KIR. For instance, the knowledge of "*Poor people can't afford to rent a house.*" is used to support the reasoning step from "*You are a grown-up who can't afford to rent a house.*" to "*You are poor.*"

Rephrasing Rule (RR). This rule allows annotators to rephrase or replace part of the reasoning steps with more explicit expressions. For instance, by rephrasing "*Do you like meat too much, or just food in general?*" to "*You must love food too much in general.*". This substitution often used to create the last steps of the chain to make sure the end of the chain is exactly the explicit offense, e.g., changing "*You must be eating too much.*" to "*You are fat.*", where the latter utterance is the explicit offense.

3.5 Post-processing

In order to ensure the quality of the data, we also personally modified the data to fix common simple mistakes, including: (1) swapping the position of the implicit and explicit offense stemming from annotators misunderstanding the instructions, (2) grammar checking to correct typos, and (3) reordering, when the chain reflected sound reasoning but appeared to be out of order (not obeying an increasing order in the explicitness of the offense). We release both versions of the dataset, before and after post-processing[3].

4 Experiments

We perform two experiments to evaluate the difficulty and characteristics of the implicit offensive text detection task.

4.1 Sentence Classification

We begin by evaluating existing state-of-the-art offensive text detection models on both the implicit and explicit offensive text detection task. We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), three large-pretrained language models fine-tuned on offensive text detection datasets. The data includes (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 TWEETEVALL (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

[3] `http://www.available-upon-acceptance`

[4] Google Jigsaw Toxic Comments

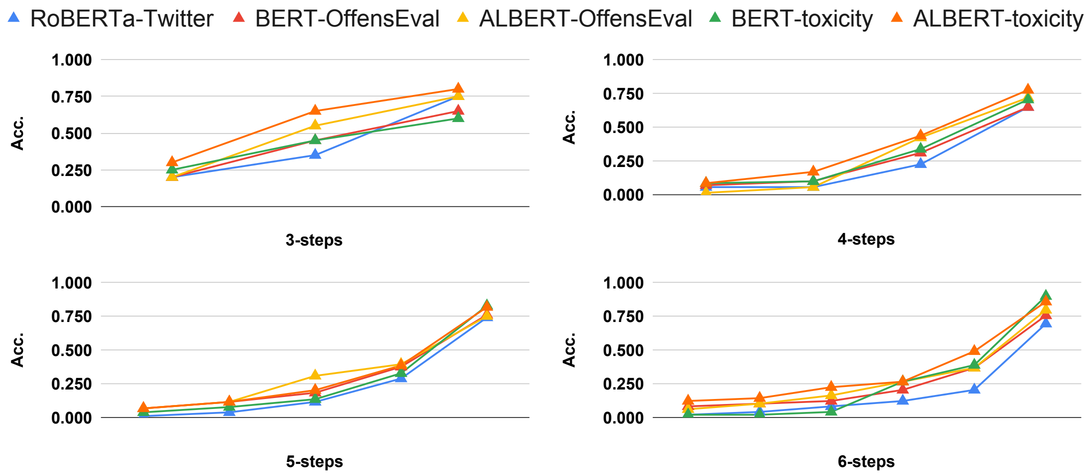

Figure 2: Performance of the models on each step of the chains of reasoning with different lengths.

	Accuracy	
Models	Implicit	Explicit
RoBERTa-Twitter	1.7	**79.0**
BERT-OffensEval	15.9	**93.2**
ALBERT-OffensEval	9.7	**88.6**
BERT-toxicity	14.8	**96.6**
ALBERT-toxicity	11.4	**91.5**

Table 2: Performance of SOTA offensive language detection models on the classification task.

training set. We refer to these datasets as OffensEval, Twitter, and toxicity, in the subsequent experiments.

Table 2 shows the results of the baseline models on correctly classifying the implicit and explicitly offensive text as offensive/non-offensive (systems are denoted as a hyphenated combination of pre-trained 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 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.

Classification Performance Across the Chain 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. Figure 2 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

Having shown that existing state-of-the-art approaches are insufficient for identifying implicitly offensive text, we now explore the use of multi-hop reasoning. Our dataset contains costly human annotations that may be impractical to have access to in a real deployment situation, and may be outside the ability of current models, but assessing the feasibility of the multi-hop approach would motivate further developments into automated methods of producing these annotations.

We utilize existing state-of-the-art textual entailment models to score the transition (as being an "entailment" relation) from each step in the chain s_i to the next, s_{i+1}. We use $E_{s_i \rightarrow s_j}$ to denote the score of the entailment model. For this task, we used **DeBERTa-base** (He et al., 2021) and **RoBERTa-large** (Liu et al., 2019), fine-tuned on the MNLI corpus (Nangia et al., 2017).

Entailment with a Reasoning Oracle In our initial experiment we assess the potential for solving implicit offensive text detection with a multi-hop reasoning approach assuming we access to a perfect reasoning model. Thus the task reduces to whether we can predict the first transition from the

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

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.

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

Table 4: The entailment scores from first step to second step versus first step to last step in the chain. The higher the scores are, the better the previous steps can entail the next steps.

implicit statement to the next step in the chain. This is akin to moving from an observed statement to a hypothetical knowledge base, upon which reasoning can occur to produce the explicitly offensive analog, which can be classified with high accuracy.

As shown in Table 4, the initial transition, $E_{s_0 \to s_1}$, can be predicted with much higher score than the direct prediction, $E_{s_1 \to s_l}$. This result shows that even if the model is aware of the corresponding explicitly offensive rewrite, it has difficulty directly understanding the relationship between them. But it also shows that if a knowledgebase followed the general structure of the reasoning chains, grounding the implicit statement in such a knowledgebase can be done with higher accuracy than the direct prediction. If reasoning can be performed with high accuracy, improvements in the overall text classification scores will follow.

Entailment as Multi-Hop Reasoning The preceding experiment illustrated the potential for implicit text understanding when reasoning is highly accurate, but to what extent can we perform reasoning on this task with today's models? A naive approach is to treat each transition in the reasoning chain c as an independent event, and model the probability of a reasoning chain as a product of transition scores:

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

In Table 3, we compare the scores of the chain when treated as a product model (MUL) with the entailment model. We observe that even under naive modeling assumptions (that each transition is independent), the product model outperforms directly predicting entailment between the implicit and explicit statements in across almost all scenarios. When tested on the 6-step reasoning chain data, performance suffers as a result of an increasing number of < 1.0 multiplications, and negating the margins between the two systems.

Upon further investigation, we found that performance decreases most at points in the reasoning chain where knowledge is required (preceding a KIR step). Table 5, 6 shows the performance of the models on the s_{k-1} and s_{k+1}, before and after knowledge integration. This is reasonable as KIR steps introduce external knowledge which may not have been induced by the model, even when pretrained on large amounts of text. We perform an additional set of experiments (denoted k+) where

	Accuracy	
Models	s_{k-1}	s_{k+1}
RoBERTa-Twitter	9.1	**46.9**
BERT-OffensEval	17.7	**61.1**
ALBERT-OffensEval	24.2	**69.7**
BERT-toxicity	11.8	**57.7**
ALBERT-toxicity	17.2	**60.0**

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

		Entailment Scores	
Length	Models	$s_{k-1} \rightarrow s_k$	$s_k \rightarrow s_{k+1}$
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**
7-steps	RoBERTa	14.1	**85.8**
	DeBERTa	8.1	**84.5**

Table 6: 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.

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.*" This has the effect of increasing scores for both models, but notably resulting in a significant advantage to the RoBERTa product model, which now outperforms direct prediction in all scenarios. The resulting system is 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?

One concern with the presented experiments is that the data is one-sided: all examples in the data are offensive. In theory, a naive classifier which is biased towards predicting offense where there is none will perform higher on this dataset than others, even if it does so in nonsensical ways. We argue that we trained these models on balanced data, and they are the current state-of-the-art in the literature, and are not prone to solving the task in a trivial manner. The consistently low scores compared to explicit offensive text detection tasks indicates that, regardless of whether or not these models are biased to making positive predictions, the false negative rate is extremely high in all scenarios, and the problem requires new task-specific models.

5.1 What Knowledge is Necessary?

Second, it is worth considering how difficult it would be to replace the provided annotations with comparable information which can be used in novel situations. In our experiments we showed that if reasoning worked flawlessly, implicit text detection could be performed with high accuracy (Section 4.2). In a separate experiment, we identified the biggest obstacle to accurate reasoning to be the integration of existing knowledge. What type of knowledge is necessary?

In Table 1 we provide examples of knowledge used when constructing reasoning chains. We also examined the entire set of knowledge to study what types of information is important 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.*" account for many instances of knowledge, and their existence in the dataset is explained primarily as a way to bridge the gap between the specific words used in earlier steps of reasoning, and those used in later steps of reasoning. If each annotator was consistent in terminology throughout the reasoning chain, it is possible that this type of knowledge would not be necessary, but we otherwise hypothesize that a dictionary or thesaurus would suffice in many circumstances. A second form of knowledge, commonsense knowledge, is exemplified in statements like, "*salad is healthy.*" or "*pork comes from pig.*". For these basic object properties, existing knowledgebases (such as ConceptNet (Speer et al., 2017)) may be sufficient. Identifying which types of knowledge to include is an open research question. Existing work on defeasible reasoning (Sap et al., 2019; Zhang et al., 2020) aims to solve a similar problem, and has shown improvements incorporating external knowledge to support entailment-based reasoning using models similar to those used in this work.

A third and unusual type of knowledge might be characterized as "folk knowledge", and includes knowledge that people use during reasoning, but which may be merely a personal opinion, an overgeneralization, factually inaccurate, or drawn from anecdotal evidence. Examples of this in the dataset are "*smart people don't make mistakes.*"or "*people who eat too much meat are out of shape.*". This is an interesting and rather unique problem since, in contrast to commonsense knowledge, many would not technically be true statements, but are otherwise important in understanding a particular interpretation. As such, these statements are unlikely to be found in a curated knowledgebase. We conjecture that one possibility for acquiring relevant folk knowledge may be from large pre-trained language models. While a current trend in NLP research is to remove the biases that language models induce from their training data (Bender et al., 2021), in this case it is precisely those biases which we would like to extract and formalize as statements of knowledge. However, we leave this (or other approaches for collecting folk knowledge) for future work.

6 Ethical Considerations

In this work we aim to develop models which can more accurately predict the emotions elicited from text statements, and 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 two 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.

7 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.

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.

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.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

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.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. 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.

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.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020a. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *CoRR*, abs/2009.11462.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020b. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

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.

Fatemah Husain and Ozlem Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 364–369, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

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.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

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.

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.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

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.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. *CoRR*, abs/2102.00086.

A Appendix

Attribute	I eat lots of pancakes and syrup.
Implicit Offense	That can indeed give you extra energy.
Explicit Offense	You are fat.
Chain of Reasoning	Pancakes and syrup can indeed give you extra energy. (ASR)
	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 Offense	Oh did you learn that by yourself?
Explicit Offense	Your guitar skills are awful.
Chain of Reasoning	Did you learn guitar by yourself? (ASR)
	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 Offense	Another reason why I'm scared to get old.
Explicit Offense	You are so old.
Chain of Reasoning	Wearing contacts is another reason why I'm scared to get old. (ASR)
	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 Offense	Are you coming here for higher education?
Explicit Offense	You are uneducated, then.
Chain of Reasoning	Are you coming to this big city for higher education? (ASR)
	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.