Leveraging Conflicts in Social Media Posts: Unintended Offense Dataset

Che-Wei Tsai[2], **Yen-Hao Huang**[1], **Tsu-Keng Liao**[2], **Didier Fernando Salazar Estrada**[1], **Retnani Latifah**[2], **Yi-Shin Chen**[1,2]

[1]Institute of Information Systems and Applications, [2]Department of Computer Science,
National Tsing Hua University, Hsinchu, Taiwan

Correspondence: yishin@gmail.com

Abstract

In multi-person communications, conflicts often arise. Each individual may have their own perspective, which can differ. Additionally, commonly referenced offensive datasets frequently neglect contextual information and are primarily constructed with a focus on intended offenses. This study suggests that conflicts are pivotal in revealing a broader range of human interactions, including instances of unintended offensive language. This paper proposes a conflict-based data collection method to utilize inter-conflict cues in multi-person communications. By focusing on specific cue posts within conversation threads, our proposed approach effectively identifies relevant instances for analysis. Detailed analyses are provided to showcase the proposed approach, efficiently gathers data on subtly offensive content. The experimental results indicate that incorporating elements of conflict into data collection not only significantly enhances the comprehensiveness and accuracy of detecting offensive language but also enriches our understanding of conflict dynamics in digital communication.

1 Introduction

Social media platforms enable communication that transcends physical boundaries and temporal limitations, allowing people from diverse backgrounds to interact regardless of direct connections. However, this diversity can also lead to challenges. Conflicts often arise from varying interpretations of responses, result in misunderstandings or offense.

While the definitions of offensive language vary, most focus on profanity and the receiver's emotional reaction rather than the sender's intention (Caselli et al., 2020). Consequently, offensiveness is often interpreted as subjective to the receiver. However, many datasets on offensive language consider only individual texts in their construction, without taking subsequent responses into account.

Additionally, numerous studies on offensive language employ "*intended affective datasets*", using intention-related language markers such as "#hate" and "#bully" for data collection. These markers indicate the sender's subjective emotional intent, not necessarily how the receiver's interprets these messages. Consequently, recipients may not perceive the message with the intended emotional intensity. Furthermore, a message not intended to offend by the sender but perceived as offensive by the receiver constitutes an unintended offense. Examples of datasets that overlook such unintended offenses include the Waseem Dataset or known as HSHP (Hateful Symbols or Hateful People) Dataset (Waseem and Hovy, 2016), Davidson Dataset or known as AHSD (Automated Hate Speech Detection) Dataset (Davidson et al., 2017), Founta dataset or known as AYR (Are You a Racist or Am I Seeing Things) Dataset (Founta et al., 2018), Kumar/TRAC 2018 Dataset (Kumar et al., 2018), HateEval Dataset (Basile et al., 2019), Offense/OLID Dataset (Zampieri et al., 2019) and Social Bias Inference Corpus (SBIC) (Sap et al., 2020).

Unintended offensive expressions are understudied compared to the intended one. Often, these unintended expressions are not meant to harm and may even appear neutral or positive (Huckin, 2002); however, recipients can still perceive offensive implications that were not intended by the sender. The interpretation of such expressions can vary greatly depending on the context. Despite widespread acknowledgment that context influences perceived offensiveness (Breitfeller et al., 2019), existing datasets typically isolate offensive expressions from their conversation context. This approach limits the analysis of offensive language and the development of detection models, prompting on an overly simplistic understanding of how offensiveness occurs in real interactions (Menini et al., 2021). Therefore, it is crucial to incorpo-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4512–4522
November 12-16, 2024 ©2024 Association for Computational Linguistics

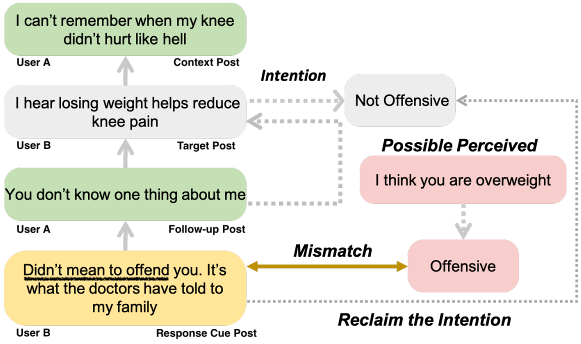

I can't remember when my knee didn't hurt like hell
User A
Context Post
I hear losing weight helps reduce knee pain
User B
Target Post
You don't know one thing about me
User A
Follow-up Post
Didn't mean to offend you. It's what the doctors have told to my family
User B
Response Cue Post
Intention
Not Offensive
Possible Perceived
I think you are overweight
Mismatch
Offensive
Reclaim the Intention

Figure 1: Inter-conflict in the conversation.

rate contextual information, such as conversation replies, when creating datasets focused on unintended and implicitly offensive language.

One method to address this is by capturing instances from the perspective of the reader, specifically through perceived emotion. Shmueli et al. (2020) introduced reactive supervision, a method to gather posts based on the perspective of a second or third person, by utilizing cue responses. These are replies that underscore the perceived emotion in a previous post. Building on this reactive supervision method, our study proposes a human-interaction-based approach for collecting offensive language.

This method extends beyond the individual who initiates the post, re-conceptualizing human interactions in light of a common phenomenon, *human conflict*. We identify two primary types of conflicts: *intra-conflict*, which represents internal inconsistency within an individual, and *inter-conflict*, which relates to perceived disagreements between individuals. Figure 1 illustrates an example of inter-conflict in a conversation, where user A takes offense at remarks made by user B, despite the lack of intended offense. This paper introduces a novel data collection method and a corresponding dataset for studying *inter-conflict*, called the Unintended Offense Dataset[1].

Exploring *human conflict* deepens our understanding of the dynamics of offensive language. By analyzing *inter-conflict* instances, we gain insights into how offensive language emerges within social interactions. Our data collection strategy captures a broader spectrum of implicit offensiveness than existing datasets and shows less topic bias. Re-evaluating human interactions through the lens of conflict allows for a more thorough exploration of the complex interplay between human expression and perception.

Our main contributions are threefold: **(1)** As far as our knowledge, this work is the **first** to focus on identifying expressions of *conflict* using innovative data collection methods for *inter-conflict* based on perceived emotion. **(2)** We developed the Unintended Offense dataset, which includes dialogue context and exhibits reduced sampling bias in topic distribution. **(3)** We demonstrate that models trained on the existing datasets are limited in their ability to identify instances of Unintended Offense.

2 Related Works

Research on offensive language has been widely conducted, including studies on dataset construction. The data collection methods for these datasets vary and include distant supervision, crowdsourcing, data aggregation, and context- or reaction-based approaches.

Distant supervision methods employ language markers—such as keywords, hashtags, emoticons, and emojis—to capture intended affective behaviors. These markers are often explicit and subjective. For instance, Waseem Dataset (Waseem and Hovy, 2016) collected the Twitter (currently X) data by searching common slurs and terms often used in religious, sexual, gender and ethnic minorities such as “MKR” and “asian drive”. The Davidson dataset (Davidson et al., 2017) utilized the Hatebase lexicon for search queries and later annotated as hate speech and offensive. Similarly, Golbeck et al. (2017) collected data using derogatory terms, hashtags, and phrases related to race and religion. A Hindi-English code-mixed offensive dataset was created using popular hashtags on sensitive topics like “beef ban” and “election result” (Kumar et al., 2018). For Evalita 2018’s Automatic Misogyny Identification task, the dataset was collected using explicit words like “bi**h” (Fersini et al., 2018), later extending to form the HatEval dataset through keyword filtering (Basile et al., 2019). The OLID dataset (Zampieri et al., 2019), collected data by searching for topic-related keywords such as “MAGA” and topic-unrelated keywords (e.g. “she is”) and was later expanded to distinguish between implicit and explicit offenses (Caselli et al., 2020).

A popular offensive public dataset by Founta et al. (2018) was created by randomly sampling Twitter data, then extracting negative sentiments and offensive words. Kumar et al. (2018) also

[1] The dataset is available at https://github.com/IDEA-NTHU-Taiwan/unintended-offense-tweets

gathered data from Facebook fanpages discussing controversial topics, alongside Twitter data. Data aggregation is another method used in offensive dataset collection, exemplified by Sap et al. (2020), where data from various existing sources and datasets like the Waseem dataset was compiled. The ISHate (Ocampo et al., 2023) dataset enhanced seven existing datasets to address implicit and subtle hate speech.

Another data collection approach involves identifying hate accounts, as done in Kwok and Wang (2013), where data was collected from self-identified or perceived racist accounts, and in ElSherief et al. (2021), which gathered tweets, retweets, and replies from three hate groups with the most followers. Some research employed crowdworkers to generate offensive language instances, resulting in implicitly abusive comparison (Wiegand et al., 2021) and euphemistic abuse (Wiegand et al., 2023) datasets. Additionally, a toxic dataset leveraging large language models has also been created (Hartvigsen et al., 2022).

The datasets mentioned earlier typically consist of standalone posts and do not encompass conversations or consider responses as contextual information. An exception is the Reddit conversation corpus augmented with automatic responses, called ToxiChat (Baheti et al., 2021). Another offensive conversational dataset is CONAN (Chung et al., 2019), which pairs hate comments with their counter-narratives. Menini et al. (2021) incorporated context by extracting previous messages from posts in the Founta dataset. However, some messages might have been deleted or could not be recovered, resulting in varied context sizes. The dataset known as the Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) (Madhu et al., 2023) is claimed to be the first to incorporate conversation context from the outset. However, while these datasets do take context into account, their collection methods primarily concentrate on intended offenses, leaving unintended understudied.

3 Unintended Offense Dataset

3.1 Response Cue

Misinterpretation in online human communication often leads to *inter-conflict*, resulting in feelings of offense among participants. Identifying instances of this *inter-conflict* allows us to recognize unintended and implicit offense. As mentioned previously, offensive language is more reflective of the receiver's reaction than the author's intention. Consequently, individuals may unintentionally offend others through their posts.

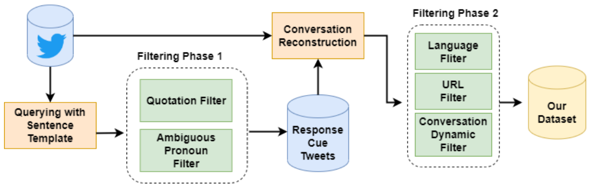


Figure 2: Overall framework

One example that we found from the wild, as illustrated in Figure 1 as mentioned previously, User A experiences offense from User B's comment: "I hear losing weight helps reduce knee pain" (highlighted in gray), and responds with, "You don't know one thing about me.". User B then follows up with, "Didn't mean to offend you" (highlighted in yellow), indicating an unawareness of the potentially offensive nature of their prior message. This scenario suggests that User B's initial comment likely contained implicit offensiveness. User B's subsequent statement, "didn't mean to offend..." signifies that any perceived offense in the previous message was unintended.

Therefore, phrases such as "Didn't mean to offend you" can function as response cues for identifying the presence of unintended offense in online communication. In order to capture the expression of implicit offense, this phrase is selected as query template for the data collection. Various dialogue contexts containing this response cue are subsequently gathered.

3.2 Proposed Framework

This paper introduce a framework to construct offensive language dataset which take into account the possible unintended and implicit offense that might happen in social media conversation. The framework can be seen in figure 2.

We leverage Twitter API 2.0 academic version to collect the cue responses by using sentence template as query. To enhance data quality and reduce noise, two specific filters are applied to the collected posts: the *quotation filter* and the *ambiguous pronoun filter*. The *quotation filter* is designed to exclude posts with a high likelihood of quoting others , while the *ambiguous pronoun filter* eliminates posts that focus on clarifying another user's intentions rather than the author's own. These filters ensure that each response cue is directly associated with the same author's prior dialogue in the thread. The dialogue contexts remaining after this filtering are referred to as *response target threads*.

Response Target Threads. *Inter-conflict* typically unfolds across several exchanges within multi-turn conversations. Our proposed method involves a comprehensive examination of these conversations, starting from the most recent cue posts and tracing back to earlier parts. This approach is predicated on the understanding that indicators of unintended, implicitly offensive content are likely to have appeared in the initial segments of the conversation.

A filtering process is applied to the *response target threads* before extracting the target offensive posts. We only included dialogue threads involving exactly two users. This allows for a more direct identification of offended users. Additionally, conversations where root posts mention other users are excluded to avoid confusion, as these often represent extensions of other dialogues.

Multi-turn Threads with Context. In scenarios where offensive threads are multi-turn, it is observed that the same author of the *cue posts* might publish multiple posts in the thread. This can lead to ambiguity regarding which post is being referred to by the cue. To resolve this, a novel *conversation dynamic filter* is introduced, designed to remove conversations where this ambiguity arises.

Definition 1. *Conversation Dynamic Filter. Given an author sequence $A^C = \{a_i^C\}$ for a dialogue thread C, the author of response cue post a_n^C is represented with X. For authors $\{a_i^C\}_{i=1}^{n-1}$ of the remaining post in the thread, each author is also represented with X if $a_i^C = a_n^C$; otherwise, a_i^C is represented by Y. A regular expression– Y+**(X)**Y+X$ – is then proposed to match the author sequence A^C represented by X and Y. If A^C mismatches the regular expression, dialogue thread C is then removed.*

To achieve this, for each thread, the post indicated by the cue post is identified as the **target post**. The posts written by the other user before the target post are regarded as the **context posts**, and the post that appears between the target post and the cue post is referred to as the **follow-up post**. Examples of this structure are illustrated in Figure 1. Furthermore, we employ a *language filter* and a *URL filter*. The *language filter* is used to ensure that only conversations in English are retained. Meanwhile, the *URL filter* is designed to exclude conversations where the target post contains URLs. This includes links to images, videos, and websites, thereby limiting the dataset to text-only expressions of offense.

Length of Threads	# of Threads	Proportion
4	3837	95.28%
5	131	3.25%
6+	59	1.47%
Total	4027	100%

Table 1: Statistics of the length of offensive threads.

3.3 Human Annotation

Searching cue posts with a query template enabled us to crawl $201k$ posts from the complete Twitter history. During the process of reconstructing conversations, we observed that posts are often missing. This is typically due to users deleting their posts or altering their privacy settings. This observation aligns with findings from previous studies (Klubicka and Fernández, 2018; Menini et al., 2021). We used the previously mentioned *Conversation Dynamic Filter* to retain conversations that matched the pattern. As a result, we successfully reconstructed approximately $42k$ conversations. However, after filtering, a total of $4,027$ fine-grained threads remained. Detailed statistics regarding the collected threads are presented in Table 1.

Among the collected data, a total of $2,401$ conversations were randomly sampled and annotated using Amazon Mechanical Turk (AMT) with each conversation receiving at least three annotations. The instruction page outlined how social media conversations may be perceived differently by individuals, covering the scope of offensiveness and the number of conversations, including a warning about adult content and advising workers to exercise discretion. It also provided examples in a template similar to the question layout. Annotators were paid approximately $\$0.40$ per assignment, with payment issued within three days.

We asked the annotators to assume themselves as the author of the context posts, this is because Hube (2020) claimed that it would increase the quality of the annotation. We also asked the annotators to assume the writer of the target posts as their unfamiliar friends, making them focus on the semantic meaning instead of inferring the relationship. The annotation schema ranged from "not offensive at all" to "extremely offensive" (range 0 - 100) and the annotators are required to provide their confidence rating with the range of 0 - 100. Conse-

Dataset	Implicitness	Collection Method	# Offensive	% Offensive
Kumar	69.40	Biased Sampling	8,716	58.10
Waseem	47.08	Biased Sampling	2,698	26.36
OLID	37.90	Biased Sampling	4,640	32.90
SBIC	54.08	Aggregation	30,593	68.31
Founta	22.13	Boosted Random Sampling	5,268	13.39
Ours $T = 50$	**74.04**	**Response-based**	1,306	54.39

Table 2: The results of the implicitness measurement.

quently, an offensiveness greater than 0 indicated the presence of offense. For annotator agreement, we calculated the average score, and if it exceeded 50, the post was considered offensive.

From the annotation process, we found that 80% (around $1,920$ conversations) of offensiveness rating were made with confidence $\geq$ 50. Meanwhile, the average offensiveness in our dataset is described by a mean of 51.71, a median of 53.66, and a standard deviation of 16.81. These figures support the method's ability to detect offensive language, as shown by the fact that more than half of the posts have an average offensiveness score above 50. Additionally, the concentration of scores near the median value of 50 indicates that most instances of language in the study are rated around this median. This aligns with the understanding that unintended implicit offensiveness, while subtle, can still be harmful.

4 Datasets Analysis

Several wide-adopted offensive datasets were compared: (1) *Kumar* (Kumar et al., 2018);; (2) *Waseem* (Waseem and Hovy, 2016); (3) *SBIC* (Sap et al., 2020); (4) *Founta* (Founta et al., 2018); and (5) *OLID* (Zampieri et al., 2019).

4.1 Implicitness

Following the procedure developed by Wiegand et al. (2019), we calculate the proportion of implicitly offensive messages among the offensive messages for each dataset above. In the existing datasets with more fine-grained classes for offensive language sub-types, the offense-related categories were consolidated into a single *offensive* category. Subsequently, the classification for measuring implicitness was simplified to binary categories: *non-offensive* and *offensive*.

For the Unintended Offense dataset, target posts with an average offensiveness score equal to or exceeding a predefined offensiveness threshold T and a confidence score of 50 or more were classified as *offensive*. Instances not meeting these criteria were labeled as *non-offensive*. Accordingly, the *lower bound* of implicitness in this dataset was determined by excluding these subtle instances.

The results of the implicitness measurement are detailed in Table 2. Among the existing datasets, *Kumar* and *SBIC* exhibited highest levels of implicitness. Notably, the implicitness in datasets obtained through biased sampling was significantly higher than the *Founta* dataset, a finding that is in line with the research of Wiegand et al. Wiegand et al. (2019). Overall, the implicitness in our dataset surpassed that of *Kumar* and *SBIC* by 5 and 13 points, respectively. This indicates that the data collection method employed in this study is more effective in capturing implicitly offensive language compared to traditional biased sampling methods.

4.2 Topic Selection Bias

Many of the offense datasets are constructed by biased sampling over the manually defined topics (Wiegand et al., 2019), and this manually defined process might introduce bias. However, as no method exists to evaluate such selection bias, comparing the bias level among datasets is challenging. To address this issue, a measurement method was proposed to reflect the level of topic selection bias by comparing dissimilarity in topics between each dataset and reference dataset.

The dataset constructed by boosted random sampling is considered to be the **reference dataset** (denoted by D^{ref}). We used Founta dataset as our reference dataset, because they have a similar topic distribution to the overall social media posts. We leveraged the topic distribution of D^{ref} to measure the degree of topic selection bias in our dataset and the dataset constructed by biased sampling. The key concept is to calculate the gap between the topic distribution of D^{ref} and the dataset to be measured, which is called the **target dataset** (denoted by D^{tar}). The gap represents how significantly the topic distribution of D^{tar} deviates from that of D^{ref}. Thus, we can regard the gap as the measurement of topic selection bias.

To model the topic distribution in datasets, a pre-trained **Latent Dirichlet Allocation** (LDA) with N topics is leveraged. For each instance in D^{ref}, a topic distribution vector $v_i^{ref} \in R^N$ is generated by LDA, where $i = 1, ..., |D^{ref}|$. The topic distribution vector of the whole reference dataset, v^{ref}, is aggregated by the following equation.

Datasets	Cosine Distance
Waseem	0.151
Kumar	0.280
Ours	0.063

Table 3: Topic selection bias comparison.

$$v^{ref} = \frac{\sum_{i=1}^{|D^{ref}|} v_i^{ref}}{|D^{ref}|} \tag{1}$$

For each instance in D^{tar}, a topic distribution vector $v_i^{tar} \in R^N$ is also generated by LDA, where $i = 1, ..., |D^{tar}|$. The topic distribution vector of the whole target dataset, v^{tar}, is aggregated by the following equation.

$$v^{tar} = \frac{\sum_{i=1}^{|D^{tar}|} v_i^{tar}}{|D^{tar}|} \tag{2}$$

The topic distribution in D^{ref} is represented by v^{ref}, and the topic distribution in D^{tar} is represented by v^{tar}. To quantify the topic selection bias in D^{tar}, we leverage the cosine distance to calculate the gap between v^{ref} and v^{tar}.

$$Bias(D^{tar}) = 1 - \frac{v^{ref} \cdot v^{tar}}{||v^{ref}|| \times ||v^{tar}||} \tag{3}$$

We leverage the proposed approach to compare the topic selection bias of our dataset with that of datasets constructed by biased sampling. The large-scale Wikipedia Corpus is utilized to train the noun-only LDA with topic number $N = 100$ to generate the coherent topics (Martin and Johnson, 2015).

To generate the topic vector representing the topic distribution of each existing dataset, the messages were solely input into the pre-trained LDA. We aggregated the outputs by Equation 1 or Equation 2 to approximate the topic distribution in each dataset. Additionally, we concatenated the context posts and target posts in our dataset before applying the pre-trained LDA. The intuition is that the context posts and target posts in the same threads would discuss the same topics. Concatenating the posts might provide more information for the LDA to model the topics.

For comparison, *Kumar* and *Waseem* were selected, since they have the highest implicitness among the biased-sampling baseline datasets. As shown in Table 3, the cosine distance of the topic distributions between our dataset and *Founta* was the smallest among all datasets. This finding shows that the topic distribution of our method is much closer to the dataset constructed by boosted random sampling (Founta). In other words, the proposed method introduces less topic selection bias than the biased sampling.

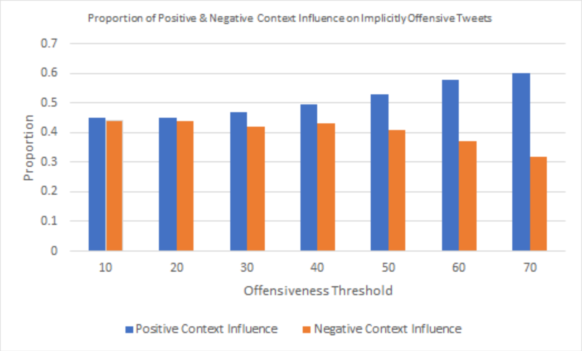


Figure 3: The proportion of positive and negative influence on implicitly offensive tweets with offensiveness $\geq T$.

Previously, Wiegand et al. (2019) found that biased sampling can be more efficient in capturing implicit offensiveness. Taking implicitness into consideration, it was discovered that the implicitness of *Kumar* is the highest of those considered, and its topic selection bias is also much higher than our dataset and *Waseem*. In other words, the measurement results showed that our data collection not only achieve the higher implicitness than existing datasets, but also introduce less topic selection bias. Thus, the proposed approach seems to alleviate the trade-off between capturing implicit offensiveness and introducing less topic selection bias, which are the limitations of boosted random sampling and biased sampling.

4.3 Context Influence

We further investigate the difference of the perceived offensiveness between situations where context is provided and those where it is absent. Using our annotation schema, we obtain offensiveness ratings both with and without context from each annotator.

Let R_{uncon} denote the rating of offensiveness without context, and R_{con} denote the rating with context. The influence of context on perceived offensiveness is denoted by δ. The calculation of context influence is formulated as follows:

$$\delta = R_{con} - R_{uncon} \tag{4}$$

A positive value of δ indicates an increase in perceived offensiveness when context is considered.

Train on	Classes	P.	R.	F1.
OLID(+,-) *N=13240*	Non-offensive	0.526	0.821	0.641
	Offensive	0.591	0.260	0.361
	Macro Avg.	0.558	0.540	0.501
Founta(+,-) *N=38872*	Non-offensive	0.526	0.954	0.678
	Offensive	0.755	0.141	0.238
	Macro Avg.	0.641	0.548	0.458
Ours (50+) **Founta (-)** *N=2088*	Non-offensive	0.732	0.969	0.834
	Offensive	0.955	0.645	0.770
	Macro Avg.	**0.843**	**0.807**	**0.802**
Ours (50+ **No-labels)** **Founta (-)** *N=5322*	Non-offensive	0.833	0.855	0.844
	Offensive	0.851	0.828	0.839
	Macro Avg.	**0.842**	**0.842**	**0.843**
Ours (all) **Founta (-)** *N=7504*	Non-offensive	0.729	0.973	0.833
	Offensive	0.960	0.637	0.766
	Macro Avg.	**0.844**	**0.805**	**0.800**

Table 4: The BERT Offense Classification Results. Test on Founta (-) and Ours (50+). *N=262/262*

Type	Example from	Classes	P.	R.	F1.
Zero Shot	-	Non-offensive	0.535	0.992	0.695
		Offensive	0.947	0.137	0.240
		Macro Avg.	0.741	0.565	0.468
One-Shot	Founta(+,-)	Non-offensive	0.560	0.985	0.714
		Offensive	0.937	0.225	0.363
		Macro Avg.	0.748	0.605	0.538

One-Shot	Ours(50+)
	Founta(-)
		Non-offensive	0.564	0.989	0.718

		Offensive	0.954	0.237	0.379
		Macro Avg.	**0.759**	**0.613**	**0.549**
5-Shot	Founta(+,-)	Non-offensive	0.568	0.950	0.711
		Offensive	0.849	0.279	0.420
		Macro Avg.	0.709	0.615	0.565

5-Shot	Ours(50+)
	Founta(-)
		Non-offensive	0.672	0.931	0.780

		Offensive	0.888	0.546	0.676
		Macro Avg.	**0.780**	**0.739**	**0.728**

Table 5: The GPT Offense Classification Results. Test on Founta (-) and Ours (50+). *N=262/262*

Conversely, a negative δ value suggests that the inclusion of context results in a decrease in perceived offensiveness.

In this study, we randomly sample two hundred conversation, each data instance is annotated by five independent crowd-workers from a pool of 72 annotators, each contributing an average of 2.78 AMT Human Intelligence Tasks (HIT). At first, annotators were shown only the target post and asked to rate its offensiveness, without considering the previous post for context. Afterward, they were given the full conversation and asked to rate the offensiveness again. The evaluation of context influence on perceived offensiveness is based on Equation 4, focusing on the differences in offensiveness ratings with and without context, as provided by the same annotator for each target post.

Figure 3 illustrates the impact of context on the perception of implicitly offensive posts under varying offensiveness thresholds T. The analysis reveals that regardless of the offensiveness levels, the implicit offensiveness is generally more prone to escalation. As T increases, there is a notable growth in the instances of positive context influence, with the disparity between positive and negative influences widening. This suggests that posts with higher offensiveness are often deemed less severe when the context is disregarded. Consequently, this emphasizes the significance of considering contextual elements in assessing the implicit offensiveness of posts.

5 Experiments

5.1 Experiments Setup

Experiments were conducted on BERT (Devlin et al., 2019) and GPT-4 (OpenAI, 2023). Since most of the collected posts were labeled as positive, two public datasets were used to construct label-balanced datasets through random sampling. These additional datasets also served as benchmarks. Since the two public datasets do not include context, we omitted the context post during the experiment and used only the target post.

In each dataset, BERT was trained for five epochs using cross-entropy loss. In the experimental setup where our dataset involved, our dataset represented the positive category, while Founta dataset was used for the negative category. The aggregated dataset was then randomly divided into training and testing sets with an 8:2 ratio, ensuring label balance. Our annotated dataset was evenly distributed across these sets, and the number of negative examples was adjusted to match the positives, ensuring a balanced representation of both categories for analysis.

For GPT-4, both zero-shot and few-shot prompting with randomly sampled examples were tested. The number of positive and negative examples shown to GPT-4 in few-shot prompts was equal. The test set consisted of our dataset as the positive category and the *Founta* dataset as the negative category.

5.2 Result and Discussion

RQ1: Can models trained on existing datasets detect Unintended Offense? The performance of single run fine-tuned BERT models is summarized in Table 4. When trained solely on the OLID and Founta datasets, these models demonstrated limited success in identifying Unintended Offense posts that were more obviously offensive, indicated by offensive values exceeding 50. The majority of posts were classified as non-offensive by these models. Notably, the model trained with OLID data showed slightly better performance, which might be attributed to the general keywords used in its data collection process.

The performance of GPT-4, as shown in Table 5, revealed that under zero-shot prompting, it faced challenges in recognizing Unintended Offense posts with higher offensive values (e.g., > 50). The introduction of positive and negative examples from the Founta dataset, in both one-shot and five-shot prompting scenarios, led to only minor improvements. In these cases, the models primarily predicted posts as non-offensive. Another experiment with GPT-4 involved introducing examples from the OLID dataset for five-shot prompting, resulting in a low macro-average F1 score of around 0.474. Similar to the results using Founta, the recall score was not high enough, indicating that Unintended Offense posts were not correctly predicted.

These experiments suggest that using existing datasets for training or introducing examples solely from existing datasets is insufficient to identify Unintended Offense posts effectively.

RQ2: Can the collected data improve Unintended Offense detection? To improve Unintended Offense detection, the study integrated newly collected data into the fine-tuning process of BERT. As shown in the results, BERT was better able to detect Unintended Offense even when trained on just $2,000$ high-quality labeled posts, although it remained more likely to predict posts as non-offensive.

Incorporating a larger dataset, consisting of $5,300$ unlabeled samples from the collection, led to a significant improvement in the macro F1 score, highlighting the utility of the collected data despite potential noise. The most notable enhancement was observed when the model was trained on the entire dataset, excluding the test split, which included less obvious Unintended Offense samples. The fully trained model successfully identified Unintended Offense, achieving a macro F1 score of 0.8.

To further investigate RQ2, the collected data were utilized as examples in few-shot prompts for GPT-4. For comparison with the performance when using only Founta data as examples, the collected data with a labeled offensive value of 50 or higher were chosen as positive examples, while Founta's negative data served as negative examples. As indicated in Table 5, the inclusion of the collected data enhanced classification performance under both one-shot and five-shot conditions. Specifically, with five-shot prompting, the collected data achieved a macro F1 score close to 0.73, demonstrating that GPT-4 could more effectively detect Unintended Offense when informed by the collected data, even with limited examples. A similar result was observed when introducing samples from the collected dataset and the OLID dataset in five-shot prompting, with the macro average F1 score close to 0.6, better than five-shot prompting with OLID dataset samples (F1 of 0.47). These findings suggest that the collected data, even in the absence of human labeling, substantially contributes to the effectiveness of Unintended Offense detection.

To explore unintended dataset detection, we analyze the classification results of unintended offensive posts that were accurately predicted as offensive when the collected dataset was introduced, but not predicted as such when it was not introduced. One example involves a context where someone is discussing an athlete from their favorite team, making the response "Lives with his Mum!" inappropriate. In another scenario, an adult mentioning drinking "Mt. Dew" receives the response "I thought only 13-year-old boys drank Mt. Dew???" This is considered offensive due to its biased assumption, even if the responder's intent was genuine curiosity. Lastly, when someone shares their daily routine, the response "OMG, are you narrating your day? Wicked" may be perceived as offensive for potentially mocking the individual.

6 Conclusion

This work proposes a data collection framework that considers the context based on post reactions. The framework captures unintended offensive tweets by leveraging the concept of *inter-conflict*. Using this approach, a dataset of Unintended Offense posts was successfully compiled. Comprehensive analyses and experiments conducted across various datasets revealed that our dataset exhibits rich emotional depth, higher implicitness, and reduced topic bias. The findings also indicate that models trained on existing datasets have difficulty to accurately recognize Unintended Offense. We believe that our *open-source* dataset and methodology will facilitate more comprehensive data analysis and research opportunities in this domain.

Limitations

In this paper, we employed human annotators to label our collected data. While some quality controls were implemented, such as requiring three annotations for each instance, subjectivity may still be present. This is partly due to the subjective nature of offensiveness, which depends on the receiver's perception and open for personal interpretation.

Furthermore, since we used Amazon Mechanical Turk, which relies on non-expert judgment, variability in the annotations might still occur. To ensure the consistency and quality of the data, only native English speakers from the USA, UK, and Australia with an approval rate higher than 90% and more than 100 approved submissions were allowed to participate. However, the selection of annotators from specific countries could lead to biases influenced by their unique backgrounds and cultural perspectives, potentially limiting the generalization of the findings. Additionally, we used the average score for annotator agreement, which overlooks any extreme cases of high and low scores within a single post, though such occurrences were rare during our annotation phase.

The models used to evaluate the datasets are limited to BERT and ChatGPT version 4.0, and the range of datasets we compared is also restricted. Additionally, the classification performed is binary, rather than a more refined classification that differentiates between implicit and explicit categories. These conditions, if considered, might result in different performance outcomes.

Ethics Statements

We collected our dataset from Twitter, complying to its policies at the time of collection. Since our dataset includes responses, we initially retained the usernames. However, after constructing the needed conversation, we replaced all usernames with @user. Despite this, the data might still retain some personal information, so responsible use of the data is encouraged.

This dataset is intended to enhance offensive language detection. Nonetheless, we acknowledge that because the dataset still contains offensive language, it is susceptible to misuse, such as generating harmful language, which is not the intention of the authors.

We used existing datasets for comparison and evaluation with our dataset. Therefore, the use of these existing datasets complied with their intended purpose, which is offensive language detection.

Acknowledgements

This research is supported by National Science and Technology Council 108-2221-E-007-064-MY3, 111-2221-E-007 -110 -MY3.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

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 Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

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.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

M Sc Christoph Hube. 2020. Methods for detecting and mitigating linguistic bias in text corpora.

Thomas N. Huckin. 2002. Critical discourse analysis and the discourse of condescension.

Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, page 16.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):1621–1622.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Fiona Martin and Mark Johnson. 2015. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 111–115.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. Reactive supervision: A new method for collecting sarcasm data. *CoRR*, abs/2009.13080.

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.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore. Association for Computational Linguistics.

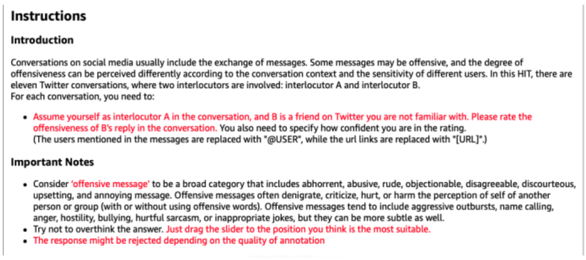
Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

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

A Appendix

A.1 Amazon Machine Turk

The details instruction of AMT is illustrated in 4, while the question layout example is shown in 5.



Instructions

Introduction

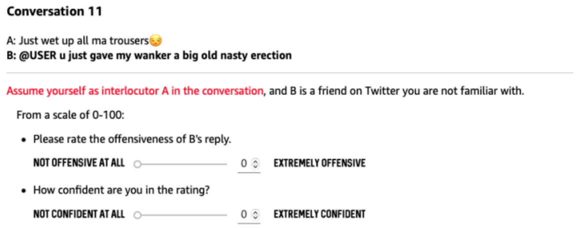
Conversations on social media usually include the exchange of messages. Some messages may be offensive, and the degree of offensiveness can be perceived differently according to the conversation context and the sensitivity of different users. In this HIT, there are eleven Twitter conversations, where two interlocutors are involved: interlocutor A and interlocutor B.
For each conversation, you need to:

- Assume yourself as interlocutor A in the conversation, and B is a friend on Twitter you are not familiar with. Please rate the offensiveness of B's reply in the conversation. You also need to specify how confident you are in the rating.
(The users mentioned in the messages are replaced with "@USER", while the url links are replaced with "[URL]".)

Important Notes

- Consider 'offensive message' to be a broad category that includes abhorrent, abusive, rude, objectionable, disagreeable, discourteous, upsetting, and annoying message. Offensive messages often denigrate, criticize, hurt, or harm the perception of self of another person or group (with or without using offensive words). Offensive messages tend to include aggressive outbursts, name calling, anger, hostility, bullying, hurtful sarcasm, or inappropriate jokes, but they can be more subtle as well.
- Try not to overthink the answer. Just drag the slider to the position you think is the most suitable.
- The response might be rejected depending on the quality of annotation

Figure 4: Annotation Instruction



Conversation 11

A: Just wet up all ma trousers😒
B: @USER u just gave my wanker a big old nasty erection

Assume yourself as interlocutor A in the conversation, and B is a friend on Twitter you are not familiar with.

From a scale of 0-100:

- Please rate the offensiveness of B's reply.

NOT OFFENSIVE AT ALL ○——— 0 EXTREMELY OFFENSIVE

- How confident are you in the rating?

NOT CONFIDENT AT ALL ○——— 0 EXTREMELY CONFIDENT

Figure 5: Question Layout

A.2 GPT-4 Prompting

In our GPT-4 experiment, we used the following prompt : “Tweet: "{first_tweet}"\n Is the tweet offensive? (Yes or No)."

The first_tweet refers to the target post. For example, “Tweet: "@user Lives with his Mum!" Is the tweet offensive?”