The Hidden Language of Harm: Examining the Role of Emojis in Harmful Online Communication and Content Moderation

Anonymous ACL submission

Abstract

Social media platforms have become central to modern communication, yet they also harbor offensive content that challenges platform safety and inclusivity. While prior research has primarily focused on textual indicators of offense, the role of emojis, ubiquitous visual elements in online discourse, remains underexplored. Emojis, despite being rarely offensive in isolation, can acquire harmful meanings through symbolic associations, sarcasm, and contextual misuse. In this work, we systematically examine emoji contributions to offensive Twitter messages, analyzing their distribution across offense categories and how users exploit emoji ambiguity. To address this, we propose an LLM-powered, multi-step moderation pipeline that selectively replaces harmful emojis while preserving the tweet's semantic intent. Human evaluations demonstrate that our approach effectively reduces offensiveness while preserving semantic integrity. Our analysis also reveals heterogeneous effects across offense types, offering nuanced insights for online communication and emoji moderation.

1 Introduction

Social media platforms host an incredibly diverse range of content, which is central to how people communicate online. However, due to varying degrees of censorship policies, platforms like Twitter often become repositories for offensive language, threatening the cohesion and safety of online communities (Davidson et al., 2019). When analyzing offensive tweets, most research has focused on textual elements—explicit slurs, abusive phrases, or implicit language that reflects social biases (Caselli et al., 2020; Zampieri et al., 2019). In response, scholars have developed various approaches, many leveraging Large Language Models (LLMs), to detect offensive content across cultural and linguistic contexts (Zhou et al., 2023a; Deng et al., 2022a). Yet, despite these efforts, one critical aspect of online communication has been largely overlooked: the role of emojis in conveying offensive messages.

Emojis, as visual symbols, are embedded in the context of communication and carry more complex semantics than individual words. On the one hand, very few emojis directly convey offensive meanings: exceptions include emojis like 🖕 (middle finger) and 💩 (pile of poop), as emojis are generally not designed with the intent to offend. On the other hand, the widespread use of emojis leads to varying interpretations. Emojis, with their symbolic representation of objects or ideas through similar shapes, can convey offensive meanings. For example, users often use the 🍑 (peach) emoji to symbolize buttocks and the 💦 (droplets) emoji to symbolize sperm. Moreover, for sentiment-related emojis, one of the key characteristics of emojis is their ability to express irony or sarcasm (Hu et al., 2017). Emojis such as 🙃 (upside-down face) and 🤣 (rolling on the floor laughing) are often used to intensify offense by conveying a sarcastic tone. Even emojis typically associated with positive sentiment, such as 😍 (smiling face with heart-eyes), can take on an offensive meaning when used in inappropriate contexts, such as sexual harassment.

Given the subtle yet potent ways in which emojis contribute to offensive communication, it is essential to systematically examine their roles within online discourse. We begin by identifying emojis frequently found in offensive tweets and analyzing how they relate to different types of offensive content. To deepen our understanding, we classify offensive tweets by category and investigate which emojis are commonly used within each category.

While content moderation has traditionally focused on text (Zampieri et al., 2019; Pitsilis et al., 2018; Husain and Uzuner, 2021), we argue that emojis present a common jailbreaking way that users exploit, either deliberately or unintentionally, to convey offensive meaning through stereotypi-

cal associations. To empower users to navigate this complex landscape, we propose an audience-oriented, mitigation-focused pipeline powered by LLMs. Rather than resorting to full-text rewriting, which can eliminate linguistic nuance, our approach performs a targeted emoji replacement. This lightweight intervention is designed to reduce the perceived offense for the viewer while preserving the semantic intent. The pipeline is implemented to identify emojis that have the potential to evoke offense, and recommend emoji surrogates that preserve the tweet's semantics. Human evaluations show that this pipeline effectively reduces offensiveness while maintaining the tweet's meaning. We also analyze its heterogeneous effects across different tweet types and examine the relationship between emoji functionality and offensiveness.

We summarize our contributions as follows:

- We explore the relationship between emojis and offensive content in online communication, examining the roles emojis play under different offensive types.
- We design and implement a multi-step LLM pipeline to better moderate offensive emojis in tweets and recommend emoji surrogates.
- We conduct a human evaluation to demonstrate the effectiveness of our pipeline and analyze its heterogeneous effects across offensive types.

2 Related Work

Our work is based on two lines of existing work: emoji functionality and offensive content detection.

Emoji Functionality and Interpretation Emojis, as prevalent visual elements, have attracted the interest of researchers. The semantics embedded in emojis extend beyond a single word token, giving their various functionalities such as expressing sentiments and irony, softening tones, and enhancing communication (Ai et al., 2017; Ge, 2019; Hu et al., 2017; Miller et al., 2016; Cramer et al., 2016). The rich meanings and diverse functionalities of emojis make them useful in various tasks, including sentiment analysis, predicting user behavior, and increasing communication (Felbo et al., 2017; Chen et al., 2018, 2019; Zhou et al., 2023b; Zhou and Ai, 2022).

Offensive Content Detection The prevalence of social networks has encouraged users to develop more flexible forms of offensive behavior. Researchers have examined patterns of offense in online communication and developed various methods to detect and mitigate offensive content in text (Davidson et al., 2019, 2017; Pitsilis et al., 2018; Poletto et al., 2021). There is increasing interest in leveraging them for the effective detection of hate speech and other hidden performance bias (Huang et al., 2023; Li et al., 2023; Zhu et al., 2023; Zhou et al., 2025). Furthermore, due to their text generation capabilities, some studies have used LLMs to augment collected datasets, thus enhancing the robustness of hate speech detection models (Xiao et al., 2024; Nghiem and Daumé III, 2024).

Beyond general offense, researchers have explored offenses of different types (Vandenbosch et al., 2015; Davidson et al., 2019; Zhong et al., 2019), as each offense type tends to target different groups. Given the strong link between offense and culture, researchers have also explored offensive content across multiple languages (Pitsilis et al., 2018; Deng et al., 2022b; Husain and Uzuner, 2021; Battistelli et al., 2020) and more increasingly diverse datasets have emerged to enhance the detection of offensive content. A few recent studies study the value of emojis as signals for offensive language detection (Kirk et al., 2021; Mubarak et al., 2023; Wiegand and Ruppenhofer, 2021). Furthermore, our research contributes to the understanding emoji-based offense in two unique ways. First, we conduct a systematic, bottom-up analysis of how a wide range of emojis contribute to offensive language. Second, building on this systematic understanding, we propose a pipeline that aims to reduce offensiveness through targeted emoji replacement, beyond the task of detection.

3 Emojis in Offensive Contexts

We begin our exploration of emoji functionality in offensive contexts by analyzing their roles and distributions. To identify offensive content, we first collected a broad, random sample of public tweets geolocated to the U.S. from January 1 to December 31, 2019, via the Twitter API[1]. We then employed a two-step approach to create a high-quality dataset of offensive tweets. Given the prohibitive cost of applying LLMs to our full one-year dataset, we first used the efficient finetuned RoBERTa model

[1] https://developer.twitter.com/en/docs/twitter-api

[2] to perform a broad initial filtering (Liu et al., 2019; Barbieri et al., 2020). The goal of this step was high recall to capture a wide range of potentially offensive content. Tweets with a predicted probability greater than 0.5 were selected, and this smaller subset was then processed by GPT-4 for a high-fidelity classification of whether the tweet contained offensive content (OpenAI, 2023). Detailed prompts can be found in Appendix A.2. Our choice of GPT-4 for this large-scale annotation task was informed by prior work demonstrating that LLMs show strong agreement with human judgments on nuanced social computing tasks, including emoji interpretation (Zhou et al., 2024b; Lyu et al., 2024).

This process resulted in 9,285 annotated offensive tweets. To ensure precise annotation, we define offensive content as posts containing unacceptable language (profanity) or targeted offenses, whether direct or veiled, including insults, threats, profane language, or swear words, following the definition used in previous work (Poletto et al., 2021; Zampieri et al., 2019). Moreover, to validate the annotation quality, one author manually reviewed a random sample of 100 tweets and confirmed 91 out of 100 as offensive tweets.

3.1 Emoji Role in Offensive Tweets

To understand how emojis function in offensive tweets, we developed a taxonomy grounded in the literature on emoji functions (Section 2) and their interaction with offensive language. We categorize emojis into four roles:

- **Offensive in itself.** The emoji alone constitutes an offense, such as 🖕 (middle finger).
- **Intensify offense.** Emojis can enhance the intensity of an offensive tweet by expressing irony or sarcasm (Weissman and Tanner, 2018), thereby amplifying its offensive nature.
- **Mitigate offense.** Emojis can also soften or adjust the tone of a tweet, reducing its offensive impact (Cramer et al., 2016; Ge, 2019).
- **Unrelated to offense.** The emoji is not directly connected to the offensive content of the tweet.

Based on the proposed taxonomy, we use GPT-4 to annotate the role of all emojis present in the collected offensive tweets. Furthermore, to validate the annotation quality, one author annotated a random sample of 100 tweets from the dataset and confirmed an agreement of 83%. The distribution of emoji roles is illustrated in Figure 1. Detailed prompts can be found in Appendix A.2.

[2]https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

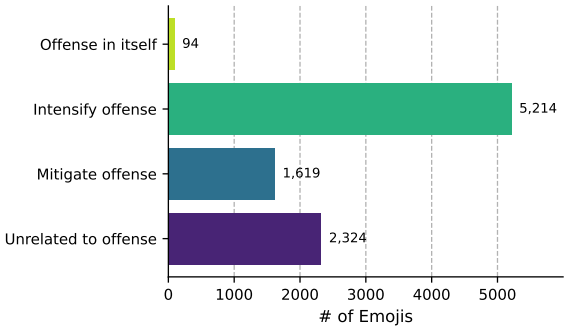

Figure 1: Distribution of emoji role in offensive tweets

Role	Top 10 Frequent Emojis
Offensive in itself	🤡💩🤮🖕😈🍆🤮🍑😡
Intensify offense	😂😭🤣🙄💀😩😫😒💯😍
Mitigate offense	😂😭👀😔🤣😅😫😩❤️😢
Unrelated to offense	😂😭😍🤣🤔🔥😫😩❤️👀

Table 1: Top-10 emojis under each emoji role in offensive tweets

The distribution (Figure 1) reveals that most annotated emojis intensify, mitigate, or are unrelated to the offense, with few being offensive in themselves, as expected given that few emojis carry inherently offensive meanings. The prevalence of intensification suggests a notable link between emojis and offensive expression. Table 1 lists the top 10 emojis for each role. Emojis categorized as "Offensive in itself" (e.g., 💩, 🖕) show no overlap with other categories and are often used for direct insults. Conversely, significant overlap exists among the top emojis for the other roles. Emojis like 😂 and 😭 appear across these categories, highlighting their context-dependent functions. Notably, even positive emojis like 😍 can intensify offense, particularly in contexts like sexual harassment.

Given that specific emojis (e.g., 🍑) seem linked to particular offensive themes, we next apply topic modeling to further explore emoji usage across different types of offensive content.

3.2 Emojis Associated with Different Offensive Topics

To better understand the offensive context in which each emoji appears, we identify the latent offensive types embedded in each tweet and explore

the emojis associated with each specific type. To summarize the offensive types across tweets, we first clustered tweets into distinct topics using unsupervised topic modeling (BerTopic (Grootendorst, 2022)). We then extracted representative words (ranked by tf-idf) for each topic and used GPT-4 to generate topic descriptions (Aizawa, 2003). We set a minimum threshold of 20 documents per cluster for our dataset, resulting in the identification of 14 distinct topics. Using the topic descriptions and representative keywords, we employed GPT-4 to summarize the offensive types and align each topic with its corresponding offense category. The types and their associated topic descriptions are presented in Table 8 in Appendix B.

To validate our GPT-4-assisted thematic grouping, we performed two checks. First, our derived categories align well with established offense types like sexual and violent offenses from related work (Vandenbosch et al., 2015; Davidson et al., 2019). Second, to quantitatively assess the rigor of the GPT-4 annotations, one author annotated a random sample of 100 tweets from the dataset and confirmed an agreement of 84%.

Based on the associated topics under each type, we further summarize the taxonomy of each offensive type, as outlined below:

- **Sexual Content and Gender Issues**: This offensive type includes sexual harassment, gender discrimination, body shaming, and objectification. Gender-based insults and derogation also fall into this category.
- **Personal Attacks and Disrespect**: This includes direct insults, disrespect, or derogation targeting individuals based on personal characteristics.
- **Racial and Ethnic Offense**: This includes racial slurs, ethnic stereotyping, and various forms of discrimination based on race or ethnicity.
- **Political and Social Issues**: This includes political attacks and harassment against individuals or groups over their political views.
- **Violence and Abuse**: This includes topics related to physical or verbal abuse and violence. This can be related to threats, aggressive behaviors, and other forms of violence as forms of offensive content.

We aggregate the emojis within each topic and use the matching relationship between topics and offensive types to assign emojis to each offensive type. In Table 2, we present the top 10 most frequent emojis for each offensive type. We note that for the type of "political and social issues" of offense, only 8 emojis are present.

Offense Type	Top 10 Frequent Emojis
Sexual Content	💦🍆👅😍🍑😈🥺🤤🔥💋
Personal Attacks and Disrespect	💩😂🤡🙄🤬🤮🐀🤣🖕😡
Racial and Ethnic Offense	😂💯😭🥴🤣💀🙄🗣️🚮😒
Political and Social Issues	🐀💩🤬😡🤡😂🖕🇺🇸
Violence and Abuse	🤬😡😤😂😠🗣️💯💀😭😒

Table 2: Top-10 emojis under different offense types. Note that there are 8 emojis for the offense type: political and social issues, in our dataset.

From Table 2, we observe that different offensive types are associated with distinct sets of frequently used emojis. The emojis used often reflect the offensive nature of the tweet. For tweets classified as "Sexual Content," we find that users frequently employ emojis such as 💦 (droplets), 🍆 (eggplant), and 👅 (tongue) to symbolize body parts. Emojis like 😍 (smiling face with heart-eyes) and 💋 (kiss), which typically convey positive sentiment, are used in these contexts to amplify the offensiveness when combined with sexual content. For the "Personal Attacks," "Racial Offense," and "Political Issues" categories, item-related emojis such as 💩 (pile of poo), 🚮 (trash), and 🐀 (rat) are commonly used to dehumanize the target and intensify the offensive content. Moreover, for the "Violence and Abuse" category, the most frequent emojis, such as 😡 (rage) and 🤬 (cursing face), reflect users' aggressive emotions and sentiments. These findings demonstrate that specific emojis are closely related to the offensive context of the tweet, amplifying the underlying offensive content.

Now that we have explored the prevalent offensive types and their associated emojis, we are also interested in understanding whether these emojis are predominantly used in offensive content or appear more frequently in unoffensive content. In the next section, we will address this question by quantifying the distribution of emojis across unoffensive and offensive tweets.

3.3 Emoji Distribution: Usage in Offensive vs. Non-Offensive Tweets

To quantify this distribution and determine whether the emojis listed in Table 1 are predominantly as-

Frequency in offensive tweets	Emojis
High frequency	🖕💩🍆🐀🤬🚾🤮👅💦🤡
Moderate frequency	🤢😠🍎😡😒😈🤬💀🗣️
Low frequency	🙄😉🤣😂😌💯🙂🥵💋😏😭😩🤔🔥😘😍

Table 3: Categorization of Emojis Based on Frequency in Offensive vs. Non-Offensive Tweets.

sociated with offensive content, we analyzed usage patterns: For each target emoji, we randomly collected a sample of 1,000 tweets containing it and calculated the proportion of offensive tweets within the sample for that specific emoji.

This percentage reflects how frequently an emoji co-occurred with offensive content within our collected samples. We then categorized the emojis based on this calculated offensive content rate:

- **High Frequency:** Emojis where over 30% of the tweets in their respective sample were offensive.
- **Moderate Frequency:** Emojis where 20% to 30% of the tweets in their respective sample were offensive.
- **Low Frequency:** Emojis where less than 20% of tweets in their respective sample were offensive.

Table 3 presents the emojis categorized according to their association level with offensive content based on this analysis.

Table 3 indicates that most emojis appearing in offensive tweets with a high frequency are item-related, used to dehumanize others or suggest body parts. For moderately frequent emojis, we find that they predominantly convey negative sentiment, such as 😤 and 😡. In contrast, emojis with positive or neutral sentiment are used less frequently in offensive tweets. This suggests that offensive tweets are more likely to be accompanied by negative or dehumanizing emojis.

4 Reduce Offense with Language Models

The analysis so far emphasize that emojis play a significant role in conveying offense, with their function closely tied to the context of the tweet.

Our next question shifts from content creation to content consumption: how can we empower users to curate a less offensive online experience? Rather than relying on platform moderation, we explore an audience-oriented intervention. This approach is motivated by recent work showing that audience desire more personalized control over their content consumption (Jhaver et al., 2023), effectively framing the goal from censoring authors to empowering viewers to manage their online experience.

While audience-side tools are promising, heavy-handed solutions like a full-text rewrite of an offensive tweet can be counterproductive. Such rewriting can drastically reduce content diversity (Padmakumar and He, 2023; Guo et al., 2023), and audience may still wish to understand the gist of the original message. Therefore, we propose a more nuanced, audience-side pipeline that acts as a lightweight filter, selectively replacing the offensive emojis. Functioning much like a browser extension, this targeted approach does not alter an author's original writing. Instead, it operates on the viewer's end to reduce their exposure to offensive content, thereby enhancing their subjective experience.

4.1 Task Setup

Our goal is to design a tool that, from the audience's perspective, replaces potentially offensive emojis with surrogates that mitigate perceived offense while preserving the semantic meaning. We design a multi-step pipeline that leverages LLMs to iteratively select appropriate emoji surrogates. This process is guided by our findings from Section 3, which we incorporate into the prompts to ensure that the replacements align with the original tweet. The input is any tweet a user encounters on their feed. The output of our pipeline is a locally rendered version of that tweet, where potentially offensive emojis have been substituted with alternatives to enhance the viewing experience.

4.2 Pipeline to Identify Emoji Surrogate

Our pipeline consists of four steps. Initially, we classify whether the tweet contains offensive content. If offensive content is detected, we then categorize the roles of each emoji into one of four categories, as described in Section 3.1. Once we identify emojis that either serve as the offense themselves or intensify the offensive tone, we leverage LLMs to suggest appropriate emoji surrogates. We present the visualization in Figure 2. Below, we provide a detailed explanation of each step and the details of prompts are shown in Appendix B.2.

Offensive Content Classification Our first step is to determine whether the tweet contains offensive content. If offensive content is identified, we

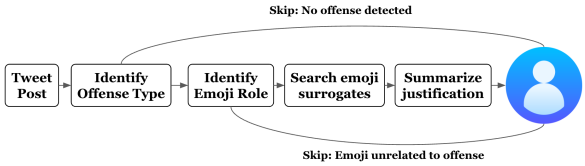

Figure 2: Multi-step pipeline to suggest the emoji surrogates for selected tweets. The blue person icon represents the end-user of our system.

prompt LLMs to classify it into one of the predefined offensive types (sexual, personal attacks, racial offense, political issues, or violence). To guide the model, we include the taxonomy of offensive types within the prompt, along with two demonstrations: one featuring an offensive tweet and the other a non-offensive tweet, to serve as examples for accurate classification. For non-offensive tweets, we leave them as is, while offensive tweets are passed to the next stage.

Determining the Role of Each Emoji We provide the LLM with the identified offense type (from step 1), our four-dimensional emoji role taxonomy and exemplars for each role. Crucially, we incorporate findings from our analysis (Section 3), such as common emojis for each offense type and role, and their general offensive frequency, into the prompt to guide the LLM towards more contextually accurate role identification.

Recommending Emoji Surrogates We ask the LLMs to suggest emoji replacements that remain consistent with the original content and sentiment of the tweet. We also include two demonstrations within the prompt. Emojis classified as "Mitigate offense" or "Unrelated to offense" are intentionally preserved to prevent losing the original semantics.

Summarize the justification Finally, we ask LLMs to summarize the reasoning of emoji replacement. The output presented a summarized justification explaining how the emoji substitutions reduce the offense level.

4.3 Experiment and Setup

For the experiment, we use GPT-4 as the LLM to recommend emoji surrogates and run the multi-step pipeline on our collected 9,285 offensive tweets (Section 3). To demonstrate effectiveness, we compare it with a **direct prompting** baseline, where the LLM is simply asked to replace emojis in offensive tweets to mitigate offense while maintaining tone (prompt details are shown in Appendix B.2).

After running our proposed pipeline, we generated emoji surrogates for a total of 7,142 tweets. For the remaining offensive tweets, no emojis were identified as playing a role in intensifying offense or directly representing offensive content.

4.4 Qualitative Evaluation

Emoji Distribution after Substitution We first examine whether offensive emojis were effectively eliminated by comparing the emoji distributions for each offensive type before and after running our pipeline and the baseline. Our analysis shows the multi-step pipeline successfully eliminates most item-based emojis frequently used for offense (e.g., 💩, 💦, 🍆) and reduces the frequency of negative sentiment emojis (😡, 🤬) and those with sarcastic tones (🙄). In contrast, the direct prompting method retained many problematic emojis (e.g., 💦, 🍑 in sexual content). This suggests our multi-step approach, informed by prior analysis, is more effective at filtering offensive emojis. (The full emoji comparison in Table 9 in Appendix B.1).

Case Studies To illustrate the pipeline's effectiveness, we present several case studies covering different offensive types. These examples showcase how the pipeline identifies the role of emojis in context and recommends appropriate, less offensive surrogates while providing step-by-step justifications. These detailed examples and the LLM's reasoning are provided in Appendix B.3.

4.5 Human Evaluation

This section details the human-centered study to assess the pipeline's effectiveness for reducing emoji-related offensiveness from a viewer's perspective.

Evaluation Design The goal of our pipeline is to reduce offensiveness while preserving semantics. A straightforward evaluation approach is to present the original and processed tweets side by side and ask audience to assess whether the pipeline effectively reduces offensive content and whether semantic meaning is preserved. However, this method may introduce cognitive bias, as audience might be inclined to perceive the two versions as inherently similar (Haselton et al., 2015).

To mitigate this bias, we conduct a within-subject user experiment in which annotators evaluate each tweet independently by answering a series of questions related to semantics and offensiveness. We recruited native English-speaking annotators with a >98% approval rate from Prolific (details in

Category	Measured Variables (*Scale / Type*)
Offensiveness	Offensiveness Score *(1-5)*, Sarcasm *(% Yes)*, Body Symbol Emojis *(% Yes)*, Dehumanizing Emojis *(% Yes)*
Semantics	Sentiment *(1-3)*, Arousal *(1-3)*, Extra Meaning *(% Yes)*, Clarity *(% Yes)*, Fluency *(% Yes)*

Table 4: Measured variables for tweet annotation. Detailed variable meanings are shown in Appendix A.3

Measured Variables	Original	Direct	Multi-step
Offensiveness Score (1-5)	3.00	2.94	2.58*
Sarcastic (% Yes)	34.0%	38.0%	37.5%
Body Symbol (% Yes)	49.5%	51.0%	52.0%
Dehumanization (% Yes)	20.0%	12.5%	16.5%
Sentiment Score (1-3)	1.48	1.59	1.64
Arousal Score (1-3)	2.14	2.06	2.01
Extra Meaning (% Yes)	25.0%	24.5%	19.0%
Clarity (% Yes)	74.0%	70.5%	72.5%
Fluency (% Yes)	77.0%	76.0%	77.0%

Table 5: Human evaluation results comparing our proposed multi-step pipeline with the direct prompting method. Statistical significance is indicated as follows: *: $p < 0.05$ (paired t-test).

Appendix A.1). Each annotator assesses 60 tweets presented in a randomized order, consisting of 20 original tweets, 20 versions of these tweets processed by our pipeline, and 20 versions processed by the baseline method. We then compare the differences in responses to the tweets before and after processing to assess the pipeline's impact. This methodology allows us to measure the change in perception for each annotator individually. This focus on the delta is powerful because it controls for the inherent subjectivity of perceiving offensiveness, providing a clearer signal of our intervention's effect. Consequently, our primary analysis relies on the statistical significance of this within-subject change, making traditional inter-annotator agreement scores a less critical measure for evaluation robustness. Ideally, annotators should perceive the rewritten tweets as less offensive.

Measures We collected measures that aim to assess the offensiveness and semantic aspect of each tweet. For offensiveness, we firstly measured the perceived level of offensiveness by asking annotators to rate how offensive they find each tweet on a scale from 1 to 5, ranging from 'not offensive' to 'extremely offensive.' Additionally, based on the findings in Sections 3.1 and 3.2, we observe that emojis can enhance offensiveness by expressing irony, symbolizing body parts, or dehumanizing the target. We ask annotators whether the emojis in each tweet exhibit these functionalities.

We also consider the influence of emoji replacement on the semantic aspect, given emojis' role, such as conveying sentiment. For each tweet, we ask annotators to assess the sentiment, emotion arousal, whether the emojis contribute external meaning, clarity, and fluency. These annotations allow us to quantify the semantic integrity of each tweet. We present the collected variables in Table 4 and the questionnaire in Appendix A.3.

Average Evaluation Result For our evaluation, we randomly sampled 600 tweets for evaluation. This sample size was determined to be sufficient for our analytical goals, as it yielded a diverse set of examples covering all identified offense types. We then present annotators with 600 tweets: 200 original tweets, 200 tweets from our multi-step pipeline, and 200 tweets via direct prompting. Each tweet is annotated by two annotators with the predefined questions. After annotation, we compute the average scores for overall offensiveness, sentiment, and arousal, as well as the percentage of 'Yes' responses for other variables. The results for the original, pipeline-processed, and direct-prompting-processed tweets are shown in Table 5.

As shown in Table 5, our proposed multi-step pipeline significantly reduces the offensiveness scores assigned by annotators. In terms of semantic preservation, tweets processed by our pipeline exhibit no notable changes in meaning. Compared to our pipeline, the direct prompting baseline achieves only a minor and statistically insignificant reduction in offensiveness. We suspect this is because, without prior knowledge of the relationship between offensiveness and emojis, LLMs struggle to identify suitable emoji surrogates.

4.6 Heterogeneous Effects by Offensive Types

While our pipeline reduces overall offensiveness, users may post offensive tweets of varying types. This raises the question of whether our pipeline's effectiveness remains consistent across offense types. We re-calculate the variables within each category and the results are presented in Table 6.

Table 6 shows our multi-step method effectively removed contextually problematic emojis. It reduced dehumanizing symbols by 12.8% in personal attacks and body-part symbols by 15.0% in sexual offenses. Corresponding semantic shifts were

Measurement Variables (Δ)	Personal Attacks		Political/Social		Racial/Ethnic		Sexual/Gender		Violence/Abuse	
	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step
Offensiveness (1-5)	−0.18	−0.05	+0.15	−0.23	−0.09	−0.94*	−0.38*	−0.38*	+0.12	−0.60*
Sarcastic (% Yes)	−2.6%	−2.6%	+12.5%*	+2.5%	+2.9%	+5.9%	+5.0%	+7.5%	0.0%	+4.3%
Body Symbol (% Yes)	+2.6%	+2.6%	+2.5%	+2.5%	+2.9%	+4.7%	0.0%	−15.0%*	+4.8%	0.0%
Dehumanization (% Yes)	−5.1%	−12.8%*	−7.5%	−22.5%*	−2.9%	0.0%	0.0%	+2.5%	0.0%	−2.4%
Sentiment (1-3)	+0.03	+0.00	+0.10	+0.25	+0.03	+0.24	+0.23	+0.13	+0.12	+0.17
Arousal (1-3)	−0.31	−0.13	−0.20	−0.38	+0.24	−0.15	−0.20	+0.00	+0.10	−0.02
Extra Meaning (% Yes)	−2.6%	−20.5%*	−2.5%	0.0%	0.0%	−2.9%	+2.5%	−7.5%	−2.4%	0.0%
Clarity (% Yes)	−2.6%	−2.6%	−10.0%	0.0%	−5.9%	0.0%	−2.5%	−2.5%	+4.8%	−2.4%
Fluency (% Yes)	+5.1%	+7.7%	−15.0%*	−5.0%	−8.8%	−8.8%	+5.0%	+2.5%	+9.5%	+2.4%

Table 6: Mean differences in human evaluation metrics after processing tweets using Direct Prompting or Multi-step Pipeline (Δ = Processed Score - Original Score), across different offense types. *: $p < 0.05$ (paired t-test).

observed: e.g., emojis conveying 'extra meaning' decreased by 20.5% alongside dehumanizing ones in personal attacks. These aligned changes suggest the pipeline correctly targeted emojis based on their function. Sarcasm levels were generally unaffected, although the direct baseline notably increased perceived sarcasm (+12.5%) while decreasing fluency (-15.0%) for political tweets, likely reflecting contextually poor emoji choices.

Crucially, removing problematic emojis did not consistently lower overall offensiveness scores, especially for highly offensive content. Despite reducing dehumanizing emojis in personal (12.8%) and political (22.5%) attacks, the change in offensiveness scores for these categories was statistically insignificant. Examining individual cases revealed that offensiveness reduction primarily occurred in mildly offensive tweets (original score < 3). A similar pattern held for sexual offenses: removing body symbols reduced offensiveness in some cases but had no impact on tweets already rated maximally offensive (score=5). This indicates that while emoji moderation effectively removes specific offensive elements, its impact on overall perceived offensiveness is limited when strong verbal attacks dominate. Emoji replacement appears most effective for mitigating milder forms of offense.

We present a case study in Table 7 in Appendix A. In the first example, replacing 💩 with (🤔, 😒) does not reduce offensiveness, as strong verbal attacks remain. In contrast, the second example shows that removing 💩 in a mildly offensive tweet lowers its perceived offensiveness. This suggests that emoji moderation is more effective in less severe cases, while text plays a dominant role in highly offensive tweets. Our analysis defines the operational boundaries of our pipeline, showing its effectiveness on mild but not severe offenses where text dominates.

5 Implications

This work examines emojis' often-overlooked role in online offensiveness. Our proposed targeted moderation pipeline offers practical tools and conceptual insights for improving online discourse. For social media users and platforms, recognizing how emojis intensify, mitigate, or reframe offense can enhance communication clarity and moderation transparency. It underscores that offense can be conveyed beyond explicit text.

For content moderation researchers, our findings emphasize the need to analyze non-textual cues. Emojis implicitly carry offense through symbolism, sarcasm, or stereotypes. We demonstrate that LLMs can selectively substitute offensive emojis while preserving semantic content.

For developers of moderation tools, this highlights the importance of incorporating non-verbal signals like emojis, moving beyond current text-centric models. It calls for broader multimodal understanding and opens opportunities for fine-tuning or prompting techniques that address text-emoji interactions in offensive communication.

6 Conclusion

In this paper, we investigate the role of emojis in offensive social media content and propose a multi-step LLM pipeline to mitigate offensiveness while preserving tweet semantics. Our analysis reveals that emojis can amplify, mitigate, or subtly alter offensive content, emphasizing the need for moderation beyond textual cues. Through human evaluation, we demonstrate that our approach effectively reduces offensiveness compared to direct prompting. Our findings highlight the importance of integrating emoji semantics into content moderation and encourage future work to explore adaptive, user-aware moderation strategies.

7 Limitations

Despite the promising results, our approach has several key limitations. First, emoji interpretation varies across individuals and cultural backgrounds (Lu et al., 2016; Zhou et al., 2024b,a). In this study, our focus on English-language tweets from U.S.-based users was a deliberate methodological choice for this foundational study. This scope allowed us to first establish that a systematic relationship exists between emoji use and offensive content and to test our pipeline's feasibility in a large, data-rich environment while minimizing cross-cultural confounding variables. Consequently, while the specific semantic mappings we found are English-centric, we argue that our core methodology, the framework for analyzing emoji roles and performing targeted replacement, is generalizable. We therefore position our work as a proof-of-concept that offers an adaptable blueprint for future work, where applying this framework across different languages and cultures remains a crucial next step.

Second, our dataset's scope is intentionally focused on a specific region (U.S.) and time period (2019). This was a deliberate choice to establish a foundational proof-of-concept using a large-scale, stable, pre-COVID baseline dataset. However, we acknowledge this limits the direct generalizability of our findings regarding specific emoji trends, as online communication evolves. While the fundamental behavioral patterns we identify (e.g., using emojis for sarcasm or dehumanization) are likely durable, the specific emojis used to express these patterns may change over time and across regions. Future work should validate these patterns on more contemporary and geographically diverse datasets.

Third, our study was designed to first answer a critical prerequisite question: can our intervention effectively reduce perceived offensiveness in a controlled environment? Our work provides this foundational validation, demonstrating that the pipeline is successful at its primary task. We acknowledge that this does not measure real-world user acceptance or behavioral responses, which is a vital next step. However, this constitutes a different type of research question that requires a distinct experimental setup (e.g., a custom user interface and a longitudinal study), and we frame our current work as an essential precursor to such future user-centered studies.

Fourth, the effectiveness of our pipeline is inherently tied to the capabilities and potential biases of the LLMs it employs (e.g., GPT-4, RoBERTa). These models, despite their advancements, can reflect biases present in their training data, potentially leading to skewed interpretations of emoji offensiveness or unfair targeting of certain emoji uses or user expressions. Furthermore, the generalization of the pipeline to novel, rapidly evolving emoji slang or newly introduced emojis is a continuous challenge. LLMs may not immediately grasp the nuanced offensive uses of emojis that emerge after their last training update, requiring ongoing monitoring and model fine-tuning. Therefore, while we justify our use of LLMs for their scalable analytical power, we emphasize that our human evaluation (Section 4.5) serves as the final arbiter of our method's success, providing a crucial check against these potential model-centric biases.

8 Ethical Consideration

Ethical considerations for the annotation process were carefully observed. The privacy of annotators was protected as no personally identifiable information was collected. The task involved evaluating tweet content and did not entail extensive or intrusive tool usage. In line with our Institutional Review Board's (IRB) protocols for research not involving the collection of identifiable private information about human subjects, this portion of the study was deemed exempt from formal IRB review. We also note that AI assistants were employed to support coding tasks during the implementation of our experiments.

References

Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *ICWSM 2017*.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. 2020. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.

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.

Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanzhe Liu. 2018. Through a gender lens. *WWW 2018*.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered representation learning for Cross-Lingual sentiment classification. In *WWW 2019*.

Henriette Cramer, Paloma de Juan, and Joel Tetreault. 2016. Sender-intended functions of emojis in us messaging. In *MobileHCI 2016*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

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

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022a. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Yong Deng, Chenxiao Dou, Liangyu Chen, Deqiang Miao, Xianghui Sun, Baochang Ma, and Xiangang Li. 2022b. Beike nlp at semeval-2022 task 4: prompt-based paragraph classification for patronizing and condescending language detection. *arXiv preprint arXiv:2208.01312*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *EMNLP*.

Jing Ge. 2019. Emoji sequence use in enacting personal identity. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 426–438, New York, NY, USA. Association for Computing Machinery.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Martie G Haselton, Daniel Nettle, and Paul W Andrews. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Nguyen, and Jiebo Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM 2017*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

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. *arXiv preprint*.

Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. In *UbiComp 2016*.

Hanjia Lyu, Weihong Qi, Zhongyu Wei, and Jiebo Luo. 2024. Human vs. lmms: Exploring the discrepancy in emoji interpretation and usage in digital communication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2104–2110.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "blissfully happy" or "ready tofight": Varying interpretations of emoji. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 259–268.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457.

Huy Nghiem and Hal Daumé III. 2024. Hatecot: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. *arXiv preprint arXiv:2403.11456*.

OpenAI. 2023. Gpt-4 technical report.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

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.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Laura Vandenbosch, Johanna MF van Oosten, and Jochen Peter. 2015. The relationship between sexual content on mass media and social media: A longitudinal study. *Cyberpsychology, Behavior, and Social Networking*, 18(12):697–703.

Benjamin Weissman and Darren Tanner. 2018. A strong wink between verbal and emoji-based irony: How the brain processes ironic emojis during language comprehension. *PloS one*, 13(8):e0201727.

Michael Wiegand and Josef Ruppenhofer. 2021. Exploiting emojis for abusive language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 369–380.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations. *arXiv preprint arXiv:2406.12223*.

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. *arXiv preprint arXiv:1902.09666*.

Ruiqi Zhong, Yanda Chen, Desmond Patton, Charlotte Selous, and Kathy McKeown. 2019. Detecting and reducing bias in a high stakes domain. *arXiv preprint arXiv:1908.11474*.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

Yuhang Zhou and Wei Ai. 2022. #emoji: A study on the association between emojis and hashtags on twitter. In *ICWSM 2022*.

Case	Original Tweet	Revised Tweet
No Change in Offensiveness (Score: 4)	You all are dog 💩 💩Trump is a disgrace, a traitor and must be removed from office.	You all are dog 🤔😒 Trump is a disgrace,a traitor and must be removed from office.
Offensiveness Reduced (Score: 2 → 1)	#TRUMP2020 Landside stable genius 💩	#TRUMP2020 Landside stable genius 😒

Table 7: Case study of tweets before and after dehumanizing emoji removal.

Yuhang Zhou, Giannis Karamanolakis, Victor Soto, Anna Rumshisky, Mayank Kulkarni, Furong Huang, Wei Ai, and Jianhua Lu. 2025. Mergeme: Model merging techniques for homogeneous and heterogeneous moes. *arXiv preprint arXiv:2502.00997*.

Yuhang Zhou, Xuan Lu, and Wei Ai. 2024a. From adoption to adaption: Tracing the diffusion of new emojis on twitter. *arXiv preprint arXiv:2402.14187*.

Yuhang Zhou, Xuan Lu, Ge Gao, Qiaozhu Mei, and Wei Ai. 2023b. Emoji promotes developer participation and issue resolution on github. *arXiv preprint arXiv:2308.16360*.

Yuhang Zhou, Paiheng Xu, Xiyao Wang, Xuan Lu, Ge Gao, and Wei Ai. 2024b. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. *arXiv preprint arXiv:2402.01681*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Appendix

A Supplementary Evaluation Materials

A.1 Detailed Recruitment Process

Participant recruitment was conducted via the Prolific online platform, yielding a cohort of 20 annotators. To ensure high-quality data, we used strict filtering criteria: participants were required to be adults, identify as native English speakers, and have a high platform approval rate (>98%) with extensive prior task experience. All annotators were provided with detailed instructions and examples (see Appendix A.3) to calibrate their understanding of the offensiveness scale, and we conducted a pilot study to confirm these instructions were clear regarding emoji use and interpretation. The annotation protocol involved each individual assessing 60 tweets, a task projected to require approximately 90 minutes of engagement, with compensation provided at a rate of $18 USD per hour.

A.2 Detailed Prompts for Data Annotation

For all data annotation tasks described in Section 3, we used GPT-4 with a temperature setting of 0. This was done to ensure deterministic and reproducible outputs. The following zero-shot prompts rely on clear definitions to guide the model.

Prompt for Offensive Content Classification

This prompt was used for the final, high-fidelity classification of tweets after the initial RoBERTa filtering.

Instruction:

Analyze the following tweet and determine if it contains offensive content based on the definition provided below.

Definition of Offensive Content:

Offensive content is defined as posts containing unacceptable language (profanity) or targeted offenses, whether direct or veiled, including insults, threats, profane language, or swear words.

Output Format:

Your response must strictly follow this format:

Offensive: [Yes/No]

Justification: [A brief reason for your classification based on the definition.]

Tweet to Classify: {tweet}

Prompt for Emoji Role Annotation

This prompt was used to annotate the function of each emoji within tweets that were already identified as offensive.

Instruction: You will be given an offensive tweet containing one or more emojis. For EACH emoji in the tweet, determine its function based on the definitions of the four categories provided below.

Functionality Category Definitions:

Offensive in itself: The emoji alone constitutes an offense.

Intensify offense: The emoji enhances the intensity of the offensive tweet, for example by expressing irony or sarcasm.

Mitigate offense: The emoji softens or adjusts the tone of the tweet, reducing its offensive impact.

Unrelated to offense: The emoji is not directly connected to the offensive content of the tweet.

Output Format: For each emoji, provide your response in the following format. Repeat for all emojis present in the tweet.

Emoji: [The emoji character]

Functionality: [The name of the category from the definitions above]

Justification: [A brief reason for the assigned functionality based on the definition and the tweet's context.]

A.3 Detailed Questionnaire of Human Evaluation

Annotator Background Questions

1. **What is/are your native language(s)?**
   (Please specify in the text box below.)
   [Fill in the blank]

2. **In the past month, how often do you use social media platforms (e.g., Twitter, Reddit, Instagram, TikTok, Messenger)? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

3. **In the past month, how often do you see content that includes emojis on social media platforms? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

4. **In the past month, how often do you use emojis yourself when engaging on social media platforms? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

Guidelines for Assessing Offensiveness

Offensiveness in the context of tweets refers to content that conveys disrespect, hostility, or harm toward an individual or group. This may occur through negative intent, harmful stereotypes, or inappropriate language. Offensive content is typically characterized by its potential to cause emotional distress or harm to the dignity of those targeted.

When assessing the offensiveness of a tweet, consider the following factors:

Intent Evaluation

- Does the tweet imply a negative or derogatory attitude toward any individual or group?
- Assess whether the speaker's intent appears harmful, disrespectful, or mocking.

Outcome Evaluation

- Could the tweet cause emotional harm or be perceived as offensive by the targeted individual or group?
- Reflect on how the individuals or groups mentioned in the tweet might feel upon reading it.

Content Evaluation

- Does the tweet include harmful stereotypes, slurs, or degrading language?
- Look for both explicit and implicit elements that might convey disrespect.

Acceptability Assessment

Based on the content and context, determine whether the tweet is socially acceptable:

- **Acceptable**: It is reasonable for the speaker to make this statement.
- **Unacceptable**: The speaker should not have made this statement.

Overall Offensiveness Rating

Rate the tweet's offensiveness on a scale of 1 to 5:

- 1 (Not Offensive At All): No evidence of harmful intent, outcome, or content.
- 2 (Slightly Offensive): Minimal or ambiguous evidence of harmful content or intent.
- 3 (Moderately Offensive): Clearly problematic in content or intent but without strong emotional harm.
- 4 (Highly Offensive): Contains explicit disrespect, harm, or stereotypes; likely to cause emotional harm.
- 5 (Extremely Offensive): Extreme disrespect or hostility; causes significant harm or distress.

Examples of Assessing Offensiveness

To provide more clarity, here are a few examples to illustrate how to evaluate the offensiveness of tweets:

Example 1

Tweet: "People from [Group X] are lazy and don't want to work."

- **Intent**: Likely derogatory due to stereotyping.
- **Outcome**: Likely to cause emotional harm to Group X.
- **Content**: Contains harmful stereotypes.
- **Acceptability**: Unacceptable.
- **Overall Offensiveness Rating**: 4 (Highly Offensive)

Example 2

Tweet: "I had a terrible experience at [Restaurant Y]."

- **Intent**: Negative but not targeting an individual or group.
- **Outcome**: Unlikely to cause emotional harm.
- **Content**: No harmful language.
- **Acceptability**: Acceptable.
- **Overall Offensiveness Rating**: 1 (Not Offensive At All)

Example 3

Tweet: "This looks like something only a [Group Z] person would wear."

- **Intent**: Ambiguous; may or may not be derogatory.

- **Outcome**: Could cause mild emotional harm to Group Z.
- **Content**: Implicitly reinforces stereotypes about Group Z.
- **Acceptability**: Borderline unacceptable.
- **Overall Offensiveness Rating**: 3 (Moderately Offensive)

Tweet Evaluation Form

Instructions: Please read the tweet carefully and answer the following questions based on your interpretation. Your responses will help us understand perceptions of the tweet's content.

Tweet: [Display the tweet here]

Offensiveness Assessment

a. **Overall Offensiveness**
On a scale from 1 to 5, where 1 means "Not offensive at all" and 5 means "Extremely offensive," how offensive do you find this tweet?
   - 1 - Not offensive at all
   - 2 - Slightly offensive
   - 3 - Moderately offensive
   - 4 - Very offensive
   - 5 - Extremely offensive

b. **Offensive Elements (Optional)**
If you found the tweet offensive, please briefly mention which elements contributed to this perception (e.g., language, emojis, insinuations).

c. **Is the tweet attempting to be humorous or sarcastic?**
   - Yes
   - No

d. **Does the tweet contain any emoji to symbolize human body parts?**
   - Yes
   - No

e. **Does the tweet contain any emoji to dehumanize a target?**
   - Yes
   - No

Sentiment Analysis

a. **Overall Sentiment**
How would you rate the overall sentiment of the tweet?
   - 1 - Negative
   - 2 - Neutral
   - 3 - Positive

b. **Emotional Intensity**
How emotionally intense or stimulating is the tweet? Consider the energy, excitement, or agitation it conveys, regardless of whether the sentiment is positive or negative.
   - 1 - Low arousal
   - 2 - Moderate arousal
   - 3 - High arousal

Extra Emoji Meaning

a. **Do you agree with this statement: "The meanings of all emojis in the tweet are disclosed by the text of the tweet."**
   - 1 - Disagree
   - 2 - Agree

Tweet Attributes

a. **Clarity**
Is the tweet clearly written? Does this tweet provide enough information for the reader to understand its meaning?
   - Yes
   - No

b. **Fluency**
Does the tweet sound fluent or natural? Consider whether the tweet is easy to read and flows smoothly.
   - Yes
   - No

B Supplementary Results

B.1 Qualitative Evaluation: Emoji Distribution Comparison

We present the emoji distribution for each offensive type before and after running our pipeline in Table 9.

Offensive Type	Topics
Personal Attacks and Disrespect	0_Personal Confrontations and Profanity, 1_Explicit Content Solicitation, 3_Offensive Language Usage, 6_Casual Slang and Swearing, 9_Offensive Language and Slang Usage, 10_Casual Profanity Usage
Sexual Content and Gender Issues	4_Demeaning Language Toward Women
Racial and Ethnic Offense	2_Racial Slur Usage in Conversation, 11_Racial Discrimination and Stereotyping
Political and Social Issues	5_Criticism of Trump's Statements, 13_Offensive Political and Religious Comments
Violence and Abuse	7_Child Abuse Concerns, 8_Sleep and Fatigue, 12_Student Violence

Table 8: Categories of offensive types and the belonged topics. Numbers before each topic represents the topic number, ranked by the number of tweets in this topic.

Offense Type	Pipeline	Emoji Distribution
Sexual Content	Original Direct prompting Multi-step	💦🍆👅😍🍑😈😉😏🔥💋 😍💦😊😋🔥🍑😅🍆😈💋 😊😅🌊⭐😩🙄🤔😐😍🌸
Personal Attacks	Original Direct prompting Multi-step	💩😂🤡🤬🤮🐀🤣🖕😡 😂😒😡😅😤🤡🤔😠🙄🤣 😅😐😒🤔😕😤😡😡👎😊
Racial Offense	Original Direct prompting Multi-step	😂💯😠😔🤣💀🙄🗣️📊😒 😂💯😭😅😒😐😡🤔🤣🙄 😅😐🤔👍😉😒😐😡😒😩
Political Issues	Original Direct prompting Multi-step	🐀💩🤬😡🤡😂🖕🇺🇸 🐻🤡🤔😡😒😔👎🇺🇸💩😒 🤔😐😒😠🤡😡🇺🇸😅😔🙄
Violence and Abuse	Original Direct prompting Multi-step	🤬😡😤😂😠🗣️💯💀😒😒 😡😤😒😅🗣️😭🔥💯😂😒 😐😤😡😕😒😅🚫😒👍😒

Table 9: Top 10 emojis by offense type, comparing the emojis in the "Original" tweets against the outputs of the "Direct prompting" baseline and our "Multi-step" pipeline after substitution.

B.2 LLM Pipeline Prompting Details

This appendix provides the detailed prompts used in each step of the LLM pipeline described in Section 4.2.

Step 1: Offensive Content Classification Prompt
You are tasked with analyzing a tweet for offensive content. Determine if the tweet contains any offensive language or sentiments. Offensive content includes any form of non-acceptable language (profanity) or a targeted offense (veiled or direct), such as insults, threats, profane language, or swear words.

If offensive content is detected, identify the type of offense based on the following categories:

1. **Sexual Content and Gender Issues**: Includes sexual harassment, gender discrimination, body shaming, objectification, gender-based insults, and derogation.
2. **Personal Attacks and Disrespect**: Ranges from direct insults to subtle disrespect/derogation targeting individuals/groups based on personal characteristics.
3. **Racial and Ethnic Offense**: Includes racial slurs, ethnic stereotyping, and discrimination/prejudice based on race or ethnicity.
4. **Political and Social Issues**: Encompasses political attacks, social discrimination, harassment, and aggression over political views or social status.
5. **Violence and Abuse**: Includes topics related to physical or verbal abuse, violence, threats, and aggressive behaviors.

Your Response Should Include:

- Whether the tweet is offensive (*Yes/No*).
- If offensive, the offense category (*e.g., Sexual Content and Gender Issues*).
- A brief justification for the chosen category.

Examples:

- **Tweet:** `@sinnersworldxxx` 🚀😒💋💗 `RT to f*p with a surprise in DM #sex #horny https://t.co/VtTjihrWSP`
- **Offensive:** Yes
- **Offense Category:** Sexual Content and Gender Issues

- **Justification:** This tweet promotes sexually suggestive behavior (*e.g., "RT to f*p"*) and references adult content (*e.g., #sex, #horny*), violating standards around explicit material.

- **Tweet:** `My throat hurts. Can god give me a break` 🥰.

- **Offensive:** No

- **Justification:** This tweet expresses personal discomfort casually and contains no inappropriate or offensive language.

Tweet to classify:

- **Tweet:** `{tweet}`

Step 2: Emoji Role Determination Prompt You will be given an offensive tweet identified as type: `{offense_type}`. Analyze each emoji (e.g., 😂) within this tweet to determine its functionality. An emoji may fit multiple categories.

The functionality categories are:

1. Emoji represents the **offense itself**
2. Emoji **intensifies** the offense
3. Emoji **mitigates** the offense
4. Emoji is **not directly related** to the offense

Use the following data (emojis based on provided list) to inform your analysis:

Emojis Often Associated with Specific Offense Types:

- *Sexual Content:* 😂, 🤣, 😭, 😍, 😩, 😋, 😩, 😩, 😈, 🙄, 😏, 💦, 🥺, 👀, 👅, 🔥, 🤔, 😒, 💀, ❤️, 💋, 😡
- *Personal Attacks:* 😂, 😭, 🤣, 🙄, 🤡, 😩, 💀, 💩, 🤮, 😒, 🤔, 😡, 🤬, 😳, 😔, 🤢, 😤, 😐, 😠, 😑, 👀, 🥺, 😅, 😩, 😉, 🗣️
- *Racial/Ethnic:* 😂, 😭, 🤣, 😩, 😒, 💀, 🙄, 💯, 🗣️, 🤔, 🤢
- *Political/Social:* 😂, 😡, 🙄, 😉, 🤢, 😮, 😏, 😭
- *Violence/Abuse:* 😂, 😭, 🤣, 💀, 😡, 🙄, 🤢, 💯, 🤬, 😩, 😒, 😍, 🔥

Popular Emojis by Typical Functionality:

- *Offense Itself (1):* 🤡, 💩, 🖕, 😂, 🖕, 😡, 🤣, 🖕, 😈, 🙄, 🤮, 💫, 🖕, 🐀, 😭, 😩, 😩, 🤬, 💯
- *Intensify Offense (2):* 😂, 😭, 🤣, 🙄, 💀, 😩, 😩, 😒, 💯, 😍, 😋, 🔥, 😤, 🤢, 😡, 🤡, 😈, 🤮, 😩, 🤬, 🤔, 😏, 💦, 🗣️, 🥵
- *Mitigate Offense (3):* 😂, 😭, 🥺, 😔, 🤣, 😩, 😩, ❤️, 😢, 😳, 🙃, 🥰, 💀, 💕, 😊, 😐, 😕, 😴, 😍, 😌, 😘, 🤗, 🤪
- *Unrelated (4):* 😂, 😭, 😍, 🤣, 🤔, 🔥, 😩, 😩, ❤️, 👀, 🥺, 💀, 🥰, 😎, 😳, 🗣️, 💙, 😏, 💯, 🖤

Emoji Frequency in Offensive Tweets:

- *High Freq.:* 🖕, 💩, 🍆, 🐀, 🤬, 🚮
- *Medium Freq.:* 🤮, 👅, 💦, 🤡, 🤢, 😠, 🍑, 😡, 😒, 😈, 😤, 💀, 🗣️
- *Low Freq.:* 🙄, 😩, 🤣, 😂, 😋, 💯, 😊, 😜, 💋, 😏, 😭, 😩, 🤔, 🔥, 😘, 😍

Your Task: For each emoji in the provided tweet:

- State the emoji (e.g., 😭).
- Assign functionality category/categories (1-4).
- Provide a brief justification based on context and the data provided.

Examples:

- **Tweet:** `Y'all seriously why do her boobs look like analog sticks`😭🤣
- **Emojis:** 😭, 🤣
- **Analysis for 😭:**
  - **Functionality:** 2 (Intensifies), 4 (Unrelated)
  - **Justification:** Amplifies mockery, making the comment seem excessively funny, reinforcing disrespect. Also unrelated to the core offensive content itself.
- **Analysis for 🤣:**
  - **Functionality:** 2 (Intensifies), 4 (Unrelated)
  - **Justification:** Emphasizes the offensive joke as humorous, diminishing seriousness. Also unrelated to the core offensive content.

- **Tweet:** `just remembered some of the stupid sh*t I did last night, how do I get myself in these situations` 😂😂
- **Emojis:** 😂, 😂
- **Analysis for 😂 (first):**
  - **Functionality:** 3 (Mitigates), 4 (Unrelated)
  - **Justification:** Downplays potential offense with humor/self-deprecation, framing it as non-serious.
- **Analysis for 😂 (second):**
  - **Functionality:** 3 (Mitigates), 4 (Unrelated)
  - **Justification:** Repetition further emphasizes humor, reducing seriousness.
- **Tweet:** `F**k de calor`🥵
- **Emojis:** 🥵
- **Analysis for 🥵:**
  - **Functionality:** 2 (Intensifies)
  - **Justification:** Intensifies the expressed frustration/discomfort, aligning with the profanity to amplify sentiment intensity.
- **Tweet:** `Like and RT for more clip`💦`#cum [...] #nuses... https://t.co/gSNIDclo6s`
- **Emojis:** 💦
- **Analysis for 💦:**
  - **Functionality:** 1 (Offense Itself), 2 (Intensifies)
  - **Justification:** Often implies sexual acts/arousal, directly contributing to and intensifying the explicit content.

Tweet to analyze:

- **Tweet:** `{tweet}`
- **Emojis in Tweet:** `{emoji}`

Step 3: Emoji Surrogate Recommendation Prompt You are given an offensive tweet (type: `{offense_type}`) where emojis have been classified by functionality:

1. Emoji represents the **offense itself**
2. Emoji **intensifies** the offense
3. Emoji **mitigates** the offense
4. Emoji is **not directly related** to the offense

Your task is to recommend replacements only for emojis categorized as **1 (Offense Itself)** or **2 (Intensifies Offense)**. Emojis categorized as **3 (Mitigates)** or **4 (Unrelated)** should be kept.

For emojis needing replacement:

- Suggest a replacement emoji (e.g., 🤔) that maintains the tweet's general tone but mitigates the specific offense.
- Choose replacements consistent with the tweet's original content context.

Your Response Should Include (for each replaced emoji):

- Emojis to Replace: Emoji
- Replacement Emojis: Emoji
- Justification: Brief explanation for the replacement choice.
- (Finally) Revised Tweet: The full tweet after all necessary replacements.

Examples:

- **Tweet:** `Y'all seriously why do her boobs look like analog sticks`😭🤣
- **Emojis/Functionality:** 😭 (2, 4), 🤣 (2, 4)
- *Replacement 1:*
  - Emoji to Replace: 😭
  - Replacement Emoji: 🤔
  - Justification: Maintains playful tone but shifts from ridicule to curiosity/confusion.
- *Replacement 2:*
  - Emoji to Replace: 🤣
  - Replacement Emoji: 😊
  - Justification: Keeps tweet lighthearted without amplifying offense; signals amusement softly.
- **Revised Tweet:** `Y'all seriously why do her boobs look like analog sticks`🤔😊

- **Tweet:** `Omg yes!` ✊😩 `sex havers roll call` 😈
- **Emojis/Functionality:** ✊ (4), 😩 (2), 😈 (1, 2)
- *Replacement 1:*
  - Emojis to Replace: 😩
  - Replacement Emojis: 😅
  - Justification: Retains excitement but softens intensity; conveys playful enthusiasm without explicit connotations.
- *Replacement 2:*
  - Emojis to Replace: 😈
  - Replacement Emojis: 😜
  - Justification: Maintains mischievous tone but mitigates offense; more lighthearted/less suggestive.
- **Revised Tweet:** `Omg yes!` ✊😅 `sex havers roll call` 😜

Tweet to process:

- **Tweet:** `{tweet}`
- **Emojis in Tweet (with functionalities):** `{emoji}`

Step 4: Justification Summary Generation Prompt You will be given an original offensive tweet, its revised version where some emoji were replaced to mitigate offense, and justifications for the offense type, original emoji functionalities, and emoji replacements.

Your task is to **summarize** these justifications into a single, concise paragraph explaining *why* specific emoji were replaced.

Recall the emoji functionality categories:

1. Offense Itself
2. Intensifies Offense
3. Mitigates Offense
4. Unrelated to Offense

We only replaced emojis categorized as **1** or **2**.

Your Response Should Include:

- A summary paragraph integrating the offense type, the functionality of the replaced emojis, and the reason for their replacements.

Inputs Provided:

- Original Tweet: `{tweet}`
- Emojis in Original Tweet: `{emoji}`
- Revised Tweet: `{revised_tweet}`
- Justification of Tweet Offense Type: `{offense_type}`
- Justification of Emoji Functionality: `{emoji_func}`
- Justification of Emoji Replacement: `{emoji_replace}`

Generate Justification Summary:

Prompt of Direct Prompting The prompt for the direct prompting baseline is:
You will be given a tweet with emojis. If this tweet is offensive, try to only replace the emojis with ones that maintain the tweet's tone but mitigate the offense. If the tweet is non-offensive, provide the original tweet as the revised tweet.

B.3 Qualitative Evaluation: Case Studies

We present four random examples of the original tweet and the revised tweet after processing through our pipeline, covering different offensive types in Figure 3. In addition, we include the justifications summarized by the LLMs in the final step of our pipeline for each emoji substitution.

The examples and justifications presented in Figure 3 demonstrate that our pipeline effectively identifies offensive content, provides the reasoning behind its offensiveness, and captures the role of emojis within the tweet. For instance, in the first example, the LLM accurately interprets 🍑 as a reference to a body part, which intensifies the offense. It suggests replacing it with the flower emoji 🌸 to keep the positive sentiment while reducing the offensive nature of the tweet. Moreover, in the second example, where the post includes the word "tacos" and the 😡 emoji, our pipeline detects the implicit racial offense toward Hispanic or Latino culture. It recommends replacing 😡 with 😕 to reduce the offense while maintaining the semantics of the post. In conclusion, our pipeline effectively identifies offensive content within tweets, uncovers the relationship between emojis and offensive material, and precisely recommends emoji surrogates

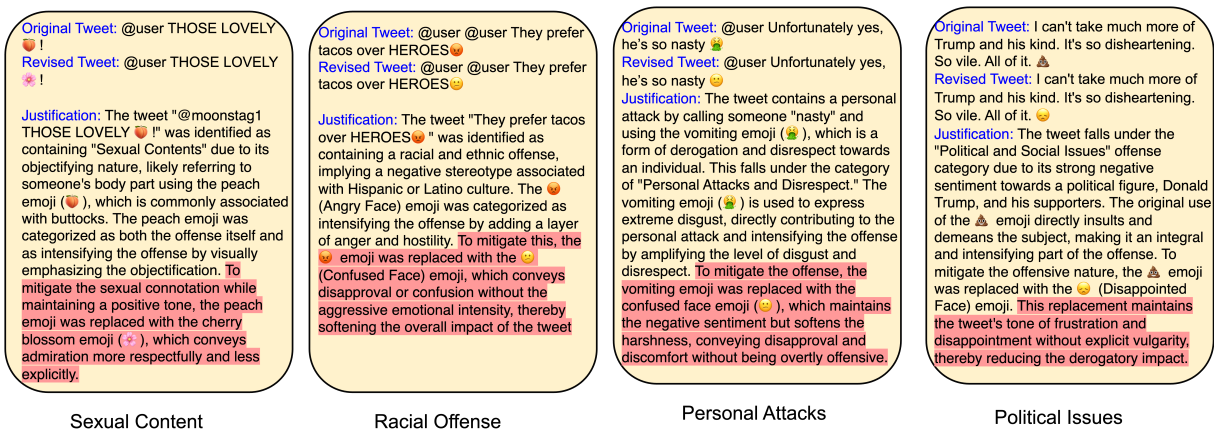

Figure 3: Justification from our multi-step pipeline of emoji replacement. The red color highlights the reason of choosing that emoji surrogate. The offense type of each tweet is labeled below.

to mitigate the offense while preserving the tweet's overall semantics.

While our case study demonstrates the effectiveness of our multi-step pipeline, the next step is to quantitatively assess its impact on offensiveness reduction for each tweet. In the following section, we leverage human annotations to evaluate the offensiveness of tweets before and after LLM rewriting.