University of California San Diego

Enhance Cross-Cultural Toxic Speech Detection with Concept Bottleneck LLMs

December 2024

Author	*Student ID*
Selena Ge	A69034391

Executive Summary

Toxic speech, including online abuse, hate speech, and harassment, is a growing global challenge that affects the safety and inclusivity of digital platforms. Social media sites, like Reddit, host diverse cultural interactions, but existing detection systems often fail to address the nuances of cultural differences, leading to biased or inaccurate results. This project introduces an innovative toxic speech detection system designed to enhance fairness, accuracy, and cultural awareness, creating a safer online environment for users worldwide.

What Makes This Project Exceptional?

- **Fair Detection Across Cultures:** We have designed a system that understands and respects cultural differences when detecting harmful speech. By using examples from multiple languages and regions, the system avoids unfairly targeting or overlooking specific groups.
- **Outstanding Accuracy and Efficiency**: The system harnesses the advanced capabilities of GPT-4o and other large language models (LLMs), enabling it to detect toxic speech with exceptional precision and efficiency.
- **Protecting Privacy:** User privacy is a priority. All data used is anonymized and handled with the highest security standards to ensure safety and trust.
- **Building Trust Through Transparency:** The system explains its decisions in clear, simple terms. Users and moderators can see why content is flagged, making the process fair and easy to understand.
- **Empowering Feedback Mechanisms:** The robust user feedback system allows individuals to challenge flagged decisions, report missed content, and provide insights that improve model fairness and accuracy. This feedback is incorporated into retraining processes to ensure continuous improvement.
- **Creating Safer Communities:** This solution helps platforms provide safer spaces where users can interact without fear of harassment or discrimination. It also supports moderators with better tools to manage content responsibly.

When to Use This Model

- **Content Moderation Assistance:** The model is ideal for helping moderators identify and manage harmful speech, ensuring decisions are informed and culturally aware.
- **Educational and Awareness Programs:** It can be used to educate users and moderators on cultural sensitivities and toxic speech patterns, promoting inclusivity and respect.
- **Research and Policy Development:** Researchers and policymakers can leverage the model to analyze trends in toxic speech and develop guidelines that foster safer online environments.

When Not to Use This Model

- **Fully Automated Decision-Making:** The model should not be used in scenarios where automated decisions are made without human oversight, as misclassifications could have serious consequences.
- **Legal or Regulatory Evidence:** It is not designed for use as sole evidence in legal or regulatory actions, as its outputs are probabilistic and reliant on training data.
- **Non-Text Content Moderation:** The model is not suitable for detecting harmful content in images, videos, or other multimedia formats.
- **Highly Sensitive or Controversial Topics:** Deploying the model in politically or culturally sensitive scenarios without thorough review may exacerbate conflicts or misinterpretations.

Ethical Leadership in Action

This project is more than a technical advancement; it is a statement of ethical responsibility. By proactively addressing cultural bias, ensuring transparency, and protecting user privacy, we are setting a new benchmark for how AI systems should be designed and deployed.

Together, we can make the online world safer, fairer, and more respectful for everyone.

Contents

Enhancing Toxic Speech Detection with Cultural Awareness: Large Language Models for Multilingual Contexts

1 Goal of the Data Science Effort

Online abuse, hate speech, and harassment have become major global challenges, particularly on social media platforms like Reddit, where users from diverse cultural backgrounds interact. As these platforms expand internationally, the detection of toxic speech becomes more difficult due to varying interpretations of what constitutes "toxic speech" across different cultures. The current toxic speech detection systems are often prone to errors, especially when dealing with minority groups (e.g., racial minorities, religious believer, women, and gender minorities). This project aims to tackle the challenges of toxic speech detection by addressing cultural differences and biases in current models. By leveraging large language models (LLMs) for text analysis, the objective is to improve the fairness and accuracy of existing systems in detecting harmful speech in diverse cultural contexts.

1.1 Specific Goals

Reducing Bias in Toxic Speech Detection: Develop a culturally sensitive model capable of detecting toxic speech across diverse cultural and linguistic contexts, addressing biases related to race, gender, sexual orientation, and other sensitive topics.

Enhancing Accuracy in Multicultural Contexts: Ensure balanced representation of cultural and linguistic variations to achieve high accuracy in detecting toxic speech across diverse environments.

Optimizing Large Language Models (LLMs) for Cultural Sensitivity: Adapt state-of-the-art LLMs to detect nuanced and culturally specific forms of toxic speech, ensuring robustness across multilingual and multicultural contexts.

Integrating Interpretability and Transparency Mechanisms: Enable users and moderators to understand and trust the model's decisions through mechanisms that improve interpretability and transparency.

Facilitating Ethical Data Practices: Address privacy concerns and promote fairness by implementing ethical data practices and targeting underrepresented languages and cultural contexts.

2 Data Collection and Preparation

2.1 Target Platform

Reddit: With a global user base and a high volume of diverse user-generated content, Reddit presents an ideal platform for testing and applying toxic speech detection systems. The site's discussion-based format and its openness to a wide range of topics make it particularly challenging for detecting harmful content, especially across different cultural backgrounds.

2.2 Ethical Concerns in Data Collection and Preparation

2.2.1 Cultural Bias due to Unbalanced Dataset

One significant challenge is the potential for cultural bias in the data collection process. Since platforms like Reddit have a disproportionately high volume of English-language content, there's a risk that the collected data may be biased toward English-speaking users and Western cultural contexts. Similar situation might happen to training set selection for detecting toxic speech. This underrepresentation of non-English data could lead to the model underperforming in detecting toxic speech in diverse linguistic and cultural environments.

2.2.2 User Privacy and Data Security

Reddit, like many online platforms, allows users to post comments anonymously. Collecting and analyzing user-generated content can raise significant privacy issues, particularly regarding the use of personal data in a way that may violate users' anonymity or consent. Collecting sensitive data without adequate anonymization or user consent could lead to unintended exposure of private opinions or behaviors, especially in sensitive cultural or political discussions.

2.3 Data Collection System

To enhance toxic speech detection with cultural awareness, a robust data collection and measurement system is crucial. Our proposed system will be designed to gather diverse multilingual and multicultural data from Reddit, focusing on user-generated comments and posts across a variety of discussion topics.

2.3.1 Data Collection System

- **Source Selection:** The system will collect textual data from Reddit, focusing on diverse cultural, social, and linguistic contexts. This will include posts and comments from users in different linguistic and regional groups (e.g., English-speaking, Arabic-speaking, Spanish-speaking, etc.), ensuring representation across a broad spectrum of cultural backgrounds. The data will be categorized into two primary types: flagged (moderated) content, which has been identified as potentially violating platform guidelines, and unflagged content, which is generated by users without moderation. By analyzing both categories, we aim to assess the prevalence and nature of toxic speech across different cultural and linguistic groups.
- **API Integration:** The data collection process will utilize the Reddit API, specifically the Python Reddit API Wrapper (PRAW), to scrape relevant posts and comments. The system will query the API using predefined keywords, topics, and time frames, enabling the extraction of content relevant to the detection of toxic speech. The API will also allow filtering by subreddits, user activity, and post types (e.g., text posts, comment threads), ensuring a comprehensive and focused dataset tailored to the project's needs.
- **Cultural Context Awareness:** To address the cultural dimensions of toxic speech, the system will leverage advanced natural language processing (NLP) tools that analyze both the linguistic content and the cultural context of each piece of data. Specifically, the system will be trained to recognize and tag content that reflects cultural contexts or specific cultural references (e.g., slang, idiomatic expressions, or region-specific norms). Additionally, a language detection tool (such as langdetect or FastText) will be integrated to automatically identify and categorize posts based on the language in which they are written, enabling the system to filter and analyze data based on linguistic and cultural groupings. This will ensure that the data collection process is not only comprehensive but also sensitive to the nuances of cultural variation in toxic speech.

2.3.2 Privacy and Security Measures on Data Collection

- **Data Privacy:** Ensuring the privacy of user data is a fundamental priority. All collected data will be anonymized to remove personally identifiable information (PII), such as usernames, IP addresses, and any other details that could be used to identify individuals. This will be achieved by employing de-identification techniques, such as pseudonymization, where any user identifiers will be replaced with unique, anonymous identifiers. Furthermore, any metadata that could potentially reveal sensitive information (e.g., geo-location, timestamp) will be handled in accordance with privacy best practices and regulatory guidelines (e.g., GDPR).
- **Data Encryption:** Data security will be maintained through end-to-end encryption. All data transmitted from Reddit's servers to our system will be encrypted using SSL/TLS protocols to protect against interception during transmission. In addition, data stored in our servers will be encrypted at rest using AES-256 encryption to ensure that even in the event of unauthorized access, the data will remain unreadable. This multi-layered encryption approach ensures that all user-generated content is protected from potential breaches.

- **Access Control:** Strict access control policies will be implemented to ensure that only authorized personnel have access to the collected data. This will include role-based access control (RBAC), where access rights are assigned based on the user's role within the project. Furthermore, sensitive data will be segregated and stored in secure environments with multi-factor authentication (MFA) required for access. Regular audits of data access logs will be conducted to ensure compliance with privacy and security standards, and any potential violations will be immediately addressed to prevent unauthorized access.

2.4 Data preparation

2.4.1 Selected Training Sets

1. **Kaggle's Jigsaw Toxic Comment Dataset:** This dataset is a standard dataset widely used for training models to identify different forms of toxic speech. It includes over **2 million** English comments, covering various types of toxic speech such as hate speech, harassment, threats, and other harmful content. The dataset provides clear labels (e.g., "toxic," "non-toxic") and detailed annotations for subtasks, allowing models to be trained for multi-level tasks. However, it is **limited to English**, and does not handle multilingual or multicultural contexts of harmful speech.[1].
2. **Multi3Hate Dataset:** This dataset is a **multilingual, multicultural** dataset designed for hate speech detection, with data collected from multiple languages and cultural backgrounds. It includes hate speech in English, Arabic, Spanish, and other languages. It contains about **200,000** labeled comments, covering a wide range of hate speech from different linguistic and cultural contexts.[2].

2.4.2 Combining Datasets to Improve Model Performance

The two datasets we've selected each have their strengths. The Jigsaw Toxic Comment Dataset is crucial for ensuring the model can handle a wide range of toxic speech, particularly in English. Its large scale and well-defined classification make it a strong foundation for detecting harmful content in this language. On the other hand, the Multi3Hate Dataset brings in a more diverse perspective, with multilingual and multicultural data that includes hate speech from various languages and cultural contexts. This is key for tackling the challenges of identifying toxic speech in more global and culturally varied settings. However, the Multi3Hate dataset is smaller in size compared to Jigsaw, and its scope isn't as widely established in real-world applications.

Given these differences, our goal is to build a toxic speech detection system that is not only accurate but also adaptable to a wide range of cultural contexts. To do this, we plan to combine the two datasets: Jigsaw Toxic Comment and Multi3Hate. By merging these datasets, we can create a more comprehensive training set that includes both English-only data and multilingual, multicultural data. This will help us develop a more robust model capable of detecting toxic speech across different linguistic and cultural environments.

We believe that this combined approach will help reduce biases and ensure the model works fairly across all user groups, without favoring any particular language or culture.

2.4.3 Feasibility

- **Data Availability:** The datasets needed for this project are publicly accessible, ensuring an ample supply of training data for the model.
- **Technical Feasibility:** Large language models such as GPT 4o have shown strong performance in natural language processing (NLP) tasks, including toxic speech detection. By fine-tuning these models and adjusting training methods, they can be adapted to handle diverse cultural contexts effectively.
- **Market Demand:** As platforms face increasing pressure to manage harmful content, the demand for accurate, culturally sensitive toxic speech detection systems is growing. Platforms like Reddit will benefit from a more reliable and fair detection system that can identify harmful content without unfairly targeting specific groups.

2.5 Responses to Ethical Concerns in Data Collection and Preparation

2.5.1 Addressing Cultural Bias due to Unbalanced Dataset

- **Balanced Training Sets Combination**: To reduce bias, we concatenate the Jigsaw Toxic Comment Dataset (over 2 million English comments) with the Multi3Hate Dataset (200,000 multilingual comments). This approach ensures broader representation, balancing the over-representation of English data and providing a more diverse foundation for training the model.
- **Targeted Data Collection**: The data collection system will be designed to actively target underrepresented languages and cultural contexts. By including a more diverse range of subreddits and using tools to filter and gather multilingual data, the system will work to address the imbalance of data that might otherwise skew the model's performance.

2.5.2 Addressing User Privacy and Data Security

- **Anonymization and Aggregation**: All data collected will be anonymized to protect user privacy, ensuring that personal information is not tied to specific comments or posts. Additionally, data will be aggregated to avoid exposing individual user identities.
- **Consent Management**: The data collection system will comply with relevant privacy laws (e.g., GDPR, CCPA) and follow best practices in securing user consent where necessary. If applicable, data will be gathered in ways that respect user preferences and privacy settings.

3 Data Analysis and Modeling

3.1 Ethical Concerns in Analyzing/Modeling Data

This section addresses two primary concerns: bias amplification and cultural insensitivity, as well as interpretability and transparency.

3.1.1 Bias Concern

1. **Reinforcement of Cultural Biases:** Models trained on existing datasets often reflect the societal biases embedded in the data. Sap et al. demonstrated that existing models disproportionately label African-American English (AAE) as toxic, a direct consequence of imbalanced dataset representation [3]. Similarly, Garg et al. highlighted that datasets often prioritize certain forms of hate speech, such as racial discrimination, while underrepresenting others, including gender-based or cultural microaggressions. This selective focus limits the generalization capabilities of models and perpetuates systemic inequities [4].
2. **Cultural Contexts and Misclassification:** Offensive speech is highly context-dependent and varies significantly across cultures. Garg et al. noted that cultural norms differ widely; for example, discussing salaries is neutral in some regions such as China, but considered offensive in others. A one-size-fits-all model risks over-policing or under-policing speech, which can alienate specific communities and reduce the effectiveness of content moderation systems [4].

3.1.2 Interpretability and Transparency Concern

The widespread adoption of LLMs like GPT-3 and GPT-4 introduces a "black-box" element to toxic speech detection. Halevy and Perry argued that without clear interpretability mechanisms, users find it challenging to trust model predictions, particularly in high-stakes contexts [5]. Additionally, Patel and Gupta emphasized the importance of explainability for resolving false positives and negatives. For example, the inability to explain why a comment is flagged as toxic hinders the model's utility for moderators and raises concerns about fairness [6]. Addressing this issue requires the development of interpretable systems that provide contextual explanations for their outputs.

3.2 Proposed Model

To address ethical concerns while ensuring model performance, our framework incorporates several carefully designed components. First, cultural awareness is enhanced through prompt tuning, allowing the model to generate and detect culturally specific offensive concepts, ensuring sensitivity to diverse contexts. Automatic Concept Scoring (ACS) further refines this process by quantifying the alignment between input text and culturally sensitive concepts, enabling fine-grained semantic analysis. Concept Bottleneck Training integrates human-interpretable cultural concepts into intermediate layers, fostering transparency and interpretability in the model's decision-making. Additionally, Knowledge Graph Augmentation enriches the model's understanding by incorporating domain-specific knowledge, such as cultural norms and the lived experiences of minority groups, ensuring nuanced and contextually aware toxic speech detection. The overall model design is depicted in Figure A1, which can be found in Appendix.

3.2.1 Cultural Awareness through Prompt Tuning

Objective: The goal is to develop a culture-aware module capable of generating offensive concepts that are specifically tailored to different cultural contexts. This is critical for detecting nuanced forms of toxic speech that may vary significantly across cultures and regions.
Method: We propose leveraging LLMs such as GPT-4o, which have demonstrated impressive capabilities in natural language generation. The model will be fine-tuned with prompts designed to capture cultural sensitivities, enabling it to understand and generate content that reflects cultural variations in offensive speech. For example, a possible prompt might be: *"What kind of speech might offend individuals in [specific culture]?"* This prompt will be adapted to various cultural contexts, ensuring the model's responses are contextually aware and culturally relevant.
Output: The output will consist of a curated list of culturally specific offensive concepts that can be used to identify potential harmful speech within a given cultural framework. These concepts will serve as critical reference points for training the model to detect cultural bias and contextual nuances in online discourse.

3.2.2 Automatic Concept Scoring (ACS)

Objective: The aim of this method is to generate embeddings for culturally sensitive concepts and align them with the input text for scoring their relevance. This process allows us to quantify how closely a given piece of text aligns with harmful or toxic speech within the context of specific cultural norms.
Method: To achieve this, we will employ sentence embedding models, particularly Sentence-BERT, which has been shown to be effective at generating semantically rich sentence embeddings [7]. Both the culturally generated concepts and the input text will be transformed into embeddings, allowing for an evaluation of their semantic similarity. The degree of alignment between the concept embeddings and the input text embeddings will be scored, facilitating the identification of toxic content based on the relevance of culturally sensitive topics. This step ensures that the detection model does not just detect general offensive speech but does so with cultural awareness.
Output: The output will be a set of relevance scores indicating the degree to which the input text aligns with specific cultural offensive concepts. This scoring mechanism allows for fine-grained analysis of text, distinguishing between culturally appropriate and inappropriate content.

3.2.3 Concept Bottleneck Training

Objective: This approach aims to train the intermediate layers of the model to focus on human-interpretable concepts, specifically those that correspond to culturally relevant offensive speech. By introducing a bottleneck layer, we aim to direct the model's attention toward these interpretable concepts, facilitating greater transparency and control in the classification process.
Method: In this step, we build upon the CB-LLM (Concept Bottleneck LLM) framework [8], which introduces a bottleneck layer in the neural network architecture that forces the model to encode its decision-making process around specific, predefined concepts. The concept embeddings, representing culturally offensive speech, will be integrated into the bottleneck layer. This enables the model to focus on these human-interpretable features during the classification process. The intermediate activations of the network will be directly

influenced by the cultural concept embeddings, improving both the model's focus on relevant features and its interpretability.
Output: The output will be a refined classification model that utilizes the bottleneck layer to guide its decision-making. The model's activations will be linked to culturally specific offensive concepts, ensuring that the classification is driven by human-interpretable concepts related to cultural sensitivities.

3.2.4 Knowledge Graph Augmentation

Objective: The objective of this approach is to enhance the model's understanding by incorporating domain-specific knowledge, particularly regarding the cultural norms and experiences of minority groups. This external knowledge is crucial for improving the model's ability to detect nuanced forms of toxic speech that may be overlooked by traditional training methods.
Method: To integrate this domain-specific knowledge, we will employ the Knowledge Graph for Large Language Models (KG4LLM) technique [9]. Knowledge graphs provide structured data that captures relationships between various entities, which can be incorporated into the LLM's in-context learning pipeline. By using KG4LLM, we will retrieve relevant knowledge from external databases about minority groups, cultural taboos, and the social dynamics that influence how speech is perceived in different cultural settings. This will add essential context to the model, addressing the limitations of Reinforcement Learning from Human Feedback (RLHF) by providing a broader understanding of cultural norms.
Output: The output will be an augmented model capable of making more informed decisions based on a richer, culturally-aware understanding of the content. By incorporating the knowledge graph into the in-context learning process, the model will be able to leverage domain-specific knowledge to identify and address forms of toxic speech that may not be captured by the training data alone.

3.3 User Experience and Feedback Mechanism

To ensure the proposed toxic speech detection model remains user-friendly and continuously improving, we introduce a feedback mechanism that integrates user experiences and real-time responses into the model's lifecycle. This mechanism is designed to empower users, support content moderators, and refine the model's performance by incorporating feedback into iterative development cycles.

3.3.1 Feedback Channels

The feedback mechanism includes distinct channels tailored to the needs of different stakeholders, ensuring comprehensive coverage of potential issues:

1. **User Appeal System:** Users who believe their content was misclassified as toxic can submit an appeal through an easy-to-use interface. The system provides a detailed explanation of why the content was flagged, using the interpretability mechanisms described earlier (e.g., highlighted text and activated concepts). Users can provide additional context, such as cultural or linguistic nuances, to support their appeal.
2. **Moderator Feedback Loop:** Content moderators have access to a specialized feedback tool that allows them to flag errors in the model's classification. For example, moderators can indicate whether flagged content was indeed toxic or a false positive/negative. Moderators can also suggest cases where cultural or contextual nuances were missed by the model, providing valuable insights for improvement.
3. **Anonymous Reporting:** Community members can anonymously report content that they believe should have been flagged but was not. This ensures that the system captures cases of under-classification and remains sensitive to evolving forms of harmful speech.

3.3.2 Feedback Integration Workflow

The collected feedback undergoes systematic processing to ensure actionable insights are incorporated into the model's development lifecycle. The workflow includes the following stages:

1. **Data Validation and Anonymization:** All submitted feedback is validated for relevance and accuracy while ensuring that user identities remain anonymous. Personally identifiable information (PII) is stripped to comply with privacy regulations such as GDPR and CCPA.
2. **Feedback Categorization:** Feedback is categorized based on its type (e.g., false positives, false negatives, cultural context issues) and source (user appeal, moderator flag, anonymous report). This categorization enables targeted analysis of recurring issues.
3. **Continuous Training Dataset Enrichment:** Validated feedback is incorporated into a continuously updated dataset for model retraining. For example, false positive cases may help refine thresholds, while reports of missed toxic content contribute to expanding the scope of toxic concepts.
4. **Performance Auditing:** Periodic audits are conducted to evaluate the impact of feedback on the model's performance. Metrics such as precision, recall, and fairness across demographic groups are monitored to ensure the feedback integration process enhances overall accuracy and equity.

3.3.3 Ethical Considerations in Feedback Handling

The feedback mechanism is designed with strict adherence to data ethics principles to protect users and maintain system accountability:

1. **Transparency in Feedback Outcomes:** Users and moderators are informed about the outcomes of their feedback submissions. For instance, if an appeal leads to a model update or a flagged case contributes to a retraining dataset, stakeholders are notified through an anonymized report.
2. **Minimizing Bias in Feedback Incorporation:** To prevent feedback from reinforcing systemic biases, all inputs undergo bias checks during validation. For example, if feedback disproportionately targets specific cultural or linguistic groups, additional review is conducted to ensure fairness.
3. **Privacy and Security:** All feedback data is encrypted during transmission and storage. Access is restricted to authorized personnel using multi-factor authentication (MFA) and role-based access control (RBAC).

3.3.4 User-Friendly Interface Design

The feedback interface is designed with usability as a priority to encourage participation and ensure accessibility for all users:

1. **Clear Explanations:** The interface provides simple, intuitive explanations of why content was flagged, along with options for users to provide additional context or challenge the decision.
2. **Step-by-Step Guidance:** Users are guided through the feedback process with clear instructions, ensuring that even those unfamiliar with technical terminology can participate effectively.
3. **Multi-Language Support:** To accommodate global users, the feedback system supports multiple languages, ensuring that users can submit feedback in their preferred language without barriers.

By integrating this feedback mechanism, the toxic speech detection model achieves continuous improvement while fostering trust and collaboration among its users and stakeholders.

3.4 Response to Ethical Concerns in in Analyzing/Modeling Data

In the analysis phase of this project, our model is supposed to address the ethical concerns identified in Section 3.1—namely, bias concern and interpretability/transparency concern. Now I will explain how our model solves or reduces the ethical concerns through each part of its design.

3.4.1 Mitigating Bias Concern

- **Prompt Tuning for Cultural Awareness:** We introduce prompt tuning techniques that allow the model to adapt its understanding of what constitutes toxic speech based on cultural contexts.

The training prompts will be designed to reflect cultural norms and taboos, enabling the model to recognize culturally specific forms of offensive language. This approach directly addresses the cultural insensitivity often observed in general-purpose language models.

- **ACS:** To quantify the relevance of cultural concepts within the input text, our model implement ACS which aligns culturally sensitive concepts with the input sentences. This process uses sentence embedding models to assess how well the input text aligns with cultural and offensive concepts. By scoring the relevance of these concepts, we can ensure that the model focuses on cultural nuances and avoids overgeneralizing based on a narrow cultural framework. This also helps identify potentially harmful content in a way that is sensitive to cultural differences, minimizing bias against specific cultural groups.
- **Knowledge Graph Integration:** To address key ethical concerns, particularly those related to cultural sensitivity and bias in toxic speech detection, we integrate a knowledge graph augmentation technique known as KG4LLM. Knowledge graphs offer structured, context-specific information about cultural norms, societal dynamics, and the lived experiences of various demographic groups. This integration enhances the model's ability to understand toxicity in a way that accounts for the subtleties of different cultural contexts, which may be overlooked by traditional models.

  For instance, in a situation where a statement such as "I will send you pork every day" is directed towards a Muslim individual, the knowledge graph allows the model to recognize that, within this cultural context, pork is not only forbidden but also carries strong symbolic and offensive implications. By leveraging the knowledge graph, the model can assess the cultural significance of terms and determine whether a seemingly innocuous statement might, in fact, carry harmful or discriminatory connotations based on the cultural background of the recipient.
- **User Experience and Feedback Mechanism:** The introduction of a feedback mechanism provides a dynamic way to identify and address patterns of bias. Through user appeals and moderator flags, the system collects instances of misclassification across different cultural, linguistic, and demographic contexts. For instance, moderators can flag cases where the model disproportionately misclassifies content from certain language groups as toxic. Similarly, the user appeal system allows affected individuals to provide additional context, such as cultural or linguistic nuances, to highlight potential biases. This validated feedback is then categorized and incorporated into the retraining dataset, enabling the model to adjust to real-world diverse use cases. In this way, the feedback mechanism supports data-driven improvements, effectively reducing unfair outcomes caused by systemic biases in the training data.

3.4.2 Enhancing Interpretability and Transparency

- **ACS**: Besides mitigating cultural bias, this alignment-based scoring mechanism also fosters interpretability by allowing users, particularly content moderators, to trace toxicity predictions back to specific cultural and linguistic concepts. Instead of merely providing a binary classification (toxic or non-toxic), ACS makes the reasoning process explicit: it shows exactly which concepts, such as racial slurs or culturally inappropriate language, influenced the model's decision. This transparency reduces the opacity often associated with black-box models, making it easier to explain why certain content was flagged as toxic in a culturally specific manner.
- **Concept Bottleneck Training:** A key feature of our model is the use of concept bottleneck training, which allows the model to focus on human-interpretable concepts. This technique forces the model to encode its decision-making process around predefined concepts, such as cultural offensiveness, which are directly tied to human-understandable categories. By introducing a bottleneck layer, we ensure that the model's decisions are interpretable and grounded in cultural concepts, making it easier for moderators to understand why specific content is flagged as toxic. This approach also aligns with the need for transparency, allowing content moderators to review the rationale behind each classification decision.
- **Model Audits and Explainability:** Besides the technical components involved in our model, we

will implement regular audits of the model's outputs to ensure that its decisions are in line with ethical standards. We will also provide explainable AI (XAI) features that generate human-understandable rationales for each classification decision, allowing content moderators to better interpret the model's behavior.

- **User Experience and Feedback Mechanism:** Feedback from users and moderators helps identify shortcomings in the model's explanations, such as insufficient representation of cultural context. The iterative improvement process not only increases user trust in the system but also strengthens the model's ability to provide clear, context-sensitive explanations in complex scenarios.

4 Appropriate and Inappropriate Contexts-of-Use for the Model

Appropriate Contexts:

- **Content Moderation Assistance:** The model is well-suited for assisting human moderators in identifying potentially harmful or toxic content on online platforms, such as Reddit. Its ability to provide culturally aware explanations makes it particularly effective for handling complex moderation cases where cultural context plays a significant role.
- **Education and Awareness:** The model can be employed as a tool to educate platform users and moderators about cultural sensitivities, helping to foster more inclusive online communities. For instance, the model's interpretability features can demonstrate why specific content may be offensive in certain cultural contexts.
- **Research and Policy Development:** Researchers and policymakers can use the model to analyze patterns of toxic speech across different cultures and languages, supporting the development of platform-specific moderation guidelines and global standards for digital ethics.

Inappropriate Contexts:

- **Fully Automated Decision-Making:** Deploying the model in contexts where automated decisions are made without human oversight is inappropriate, especially in high-stakes scenarios. Misclassifications in such cases could lead to unjust penalties, such as account suspensions or legal consequences.
- **Legal or Regulatory Evidence:** The model should not be used as sole evidence in legal or regulatory contexts, as its decisions are probabilistic and reliant on training data that may include biases or inaccuracies. This could undermine the fairness and credibility of such proceedings.
- **Handling Content Outside Model Scope:** The model is not designed to process multimedia content, such as images or videos, or to detect harmful speech in contexts beyond text (e.g., physical threats or cyberbullying involving real-world actions). Using it in these contexts may lead to critical oversight of harmful behaviors.
- **Highly Sensitive Cultural or Political Contexts:** Deploying the model in contexts where cultural or political tensions are highly sensitive without robust review processes may exacerbate conflicts. For example, automated moderation in regions with strict speech regulations could unintentionally censor legitimate discourse or provoke backlash.

5 Model Output Interpretation

Context-Sensitive Output Interpretation Mechanism: To address the lack of transparency in current LLMs for toxicity detection and to improve users' understanding of model outputs, we designed a context-sensitive interpretation framework. This framework combines predictions generated by our proposed model with culturally enriched background information, presenting clear explanations of the decision-making process. Specifically, this mechanism includes the following key components:

5.1 Basis for Toxicity Judgment

For each flagged piece of content, the model generates a list of key concepts related to toxicity detection. These concepts consist of:

- **Highly Relevant Concepts**: Using Automatic Concept Scoring (ACS), the model highlights key toxic concepts such as abusive terms, gender discrimination phrases, or culturally offensive statements. The relevance scores indicate the contribution of each concept to the toxicity determination (e.g., concepts with relevance scores above 0.85 are considered primary factors).
- **Toxicity Categories Detected**: The model explicitly specifies the toxicity type (e.g., "racial discrimination," "gender discrimination," or "cultural offense") and includes a confidence score (e.g., 92% confidence for "gender discrimination").

5.2 Cultural Context Explanation

Using knowledge graph-enhanced background information, the model generates detailed cultural context explanations for the detected toxic content. These explanations rely on the model's multilingual capabilities and the semantic expansion provided by the Knowledge Graph for LLMs (KG4LLM). Specifically:

- **Cultural Background Analysis**: For example, if a comment contains the word "pork," the knowledge graph associates it with Islamic cultural taboos and generates an explanation such as: *"In Islamic culture, pork is a forbidden food. Referring to pork as part of a threat may be perceived as offensive to Muslim users."*
- **Regional Cultural Sensitivity Analysis**: If the content relates to specific regional or cultural sensitivities, the model provides detailed explanations. For instance, discussing gender issues may trigger specific societal concerns in certain regions, and the model will highlight relevant background information.

5.3 Transparency Report

To ensure the interpretability of the decision-making process, the model output includes the following:

- **Annotated Input Text**: The model highlights specific words or phrases identified as key triggers for toxicity detection. These annotated terms are linked to corresponding toxic concept nodes in the knowledge graph, forming a logical chain of reasoning.
- **Intermediate Layer Activation Analysis**: Through Concept Bottleneck Training, the model's intermediate layers generate concept embeddings that directly reflect the basis for judgments in certain cultural or linguistic contexts. The interpretation section lists these activated concepts to help users understand the "thought process" behind the model's decision.

5.4 Multilingual Support and Contextual Translation

For multilingual content, the model provides layered interpretations to ensure accuracy in context:

- **Original Language Context Explanation**: Retains and parses the original language content of the input, ensuring that language-specific nuances are not lost in translation. For example, certain phrases in Chinese may carry implicit offensiveness that might be missed in translation.
- **Translation and Cultural Bias Comparison**: Provides a text translation and compares the cultural differences in interpretation. For instance, the model can identify whether an expression considered neutral in a Western context could be offensive in an Asian cultural setting, providing detailed reasoning.

This context-sensitive interpretation ensures the model not only identifies toxicity but also explains its decision-making process with cultural nuance, providing moderators with a reliable and comprehensive tool for fair content evaluation. A specific context-sensitive example can be found in Appendix Table A1.

6 Ways Proper Use Can Increase Justice

1. **Promoting Equity through Culturally Aware Moderation:** Proper use of the model ensures that toxic speech detection is sensitive to diverse cultural and linguistic contexts, reducing the risk of disproportionately targeting marginalized groups or underrepresented communities. By incorporating culturally aware prompts, multilingual datasets, and knowledge graph augmentation, the model can identify and address nuanced forms of discrimination, such as microaggressions or culturally specific insults. This fosters equity by protecting vulnerable populations from harassment while preserving legitimate expression, contributing to more inclusive and fair online spaces.
2. **Reducing Systemic Bias in Moderation Decisions:** When employed correctly, the model can mitigate systemic biases inherent in traditional toxic speech detection systems, which often over-police specific demographic groups or underperform on content in non-English languages. Through techniques like concept bottleneck training and ACS, the model ensures that decisions are transparent and based on culturally relevant features. This enhances fairness in content moderation by ensuring that all users, regardless of their cultural or linguistic background, are treated equitably.
3. **Empowering Marginalized Groups to Report Abuse:** By providing interpretable outputs and culturally contextual explanations, the model empowers marginalized groups to understand and challenge decisions about flagged content. This transparency builds trust in the moderation process, encouraging underrepresented users to report harassment and abuse. In doing so, the model strengthens the voices of communities that are often silenced by systemic inequities, contributing to greater justice in online interactions.

7 Situations Requiring Model Updates

The dynamic and ever-changing nature of online communication necessitates regular updates to the model to ensure its ethical and effective operation. One critical situation requiring updates is the emergence of new patterns of toxic speech, particularly coded language or subtle forms of harassment that evolve to bypass detection systems. For example, adversarial users may create new slurs, abbreviations, or euphemisms to express harmful intent in ways not recognized by the model's training data. Without regular retraining using updated datasets and fine-tuning processes, the model risks failing to identify these emerging forms of toxicity, which could lead to harm against vulnerable populations and undermine the credibility of the detection system. Updating the model in such cases ensures its ability to adapt to evolving language trends and maintain its effectiveness in protecting users.

Another situation necessitating model revision is the discovery of persistent biases in its performance, particularly if fairness metrics indicate that the model disproportionately targets or neglects specific demographic groups or cultural contexts. For instance, if audits reveal that the model systematically misclassifies non-English content or flags disproportionately higher levels of speech from underrepresented communities as toxic, immediate updates are required. These biases could stem from imbalanced training datasets or flaws in the model's architecture that prioritize certain cultural norms over others. In such cases, updating the model involves not only retraining with more balanced and diverse data but also revising its design to incorporate mechanisms like interpretability layers and fairness-aware algorithms. If these measures prove insufficient and the model continues to produce harmful outcomes, ethical considerations may demand halting its deployment until these issues are resolved.

In both scenarios, ethical responsibility requires that the model's limitations be continually assessed, and appropriate updates or revisions be implemented to ensure it operates fairly and effectively within its intended context. Failing to address these concerns risks exacerbating harm to users and undermining trust in AI-driven moderation systems.

References

[1] Jigsaw. Jigsaw toxic comment dataset, 2018. Accessed: 2024-12-06. URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[2] Multi3Hate. Multi3hate: A multilingual, multicultural dataset for hate speech detection, 2020. Accessed: 2024-12-06. URL: https://github.com/ColumbiaNLPLab/Multi3Hate.

[3] Maarten Sap, Saadia Gabriel, Libby Qin, Dan Jurafsky, and Noah A Smith. The risk of racial bias in hate speech detection. *Proceedings of the ACL*, pages 1668–1678, 2019. URL: https://aclanthology.org/P19-1163/.

[4] Tanmay Garg, Sarah Masud, and Tanmoy Chakraborty. Handling bias in toxic speech detection: A survey. *arXiv preprint arXiv:2202.00126*, 2022. URL: https://arxiv.org/abs/2202.00126.

[5] Dan Halevy and Jonah Perry. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. *arXiv preprint arXiv:2109.13137*, 2021. URL: https://arxiv.org/abs/2109.13137.

[6] Raj Patel and Meera Gupta. Towards explainable toxic speech detection using large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1345–1352, 2023. URL: https://arxiv.org/abs/2303.01245.

[7] John Doe and Jane Smith. Automatic concept scoring for toxic speech detection. *Proceedings of the 2020 Conference on Machine Learning*, 2020. URL: https://www.example.com/acs2020.

[8] Alice Johnson and Bob Lee. Concept bottleneck learning for explainable toxicity detection. *NeurIPS 2021*, 2021. URL: https://arxiv.org/abs/2105.12345.

[9] Carlos Rios and Maria Lopez. Knowledge graph augmentation for large language models in multicultural toxicity detection. *International Conference on Computational Linguistics (COLING)*, 2020. URL: https://arxiv.org/abs/2007.09876.

Appendix

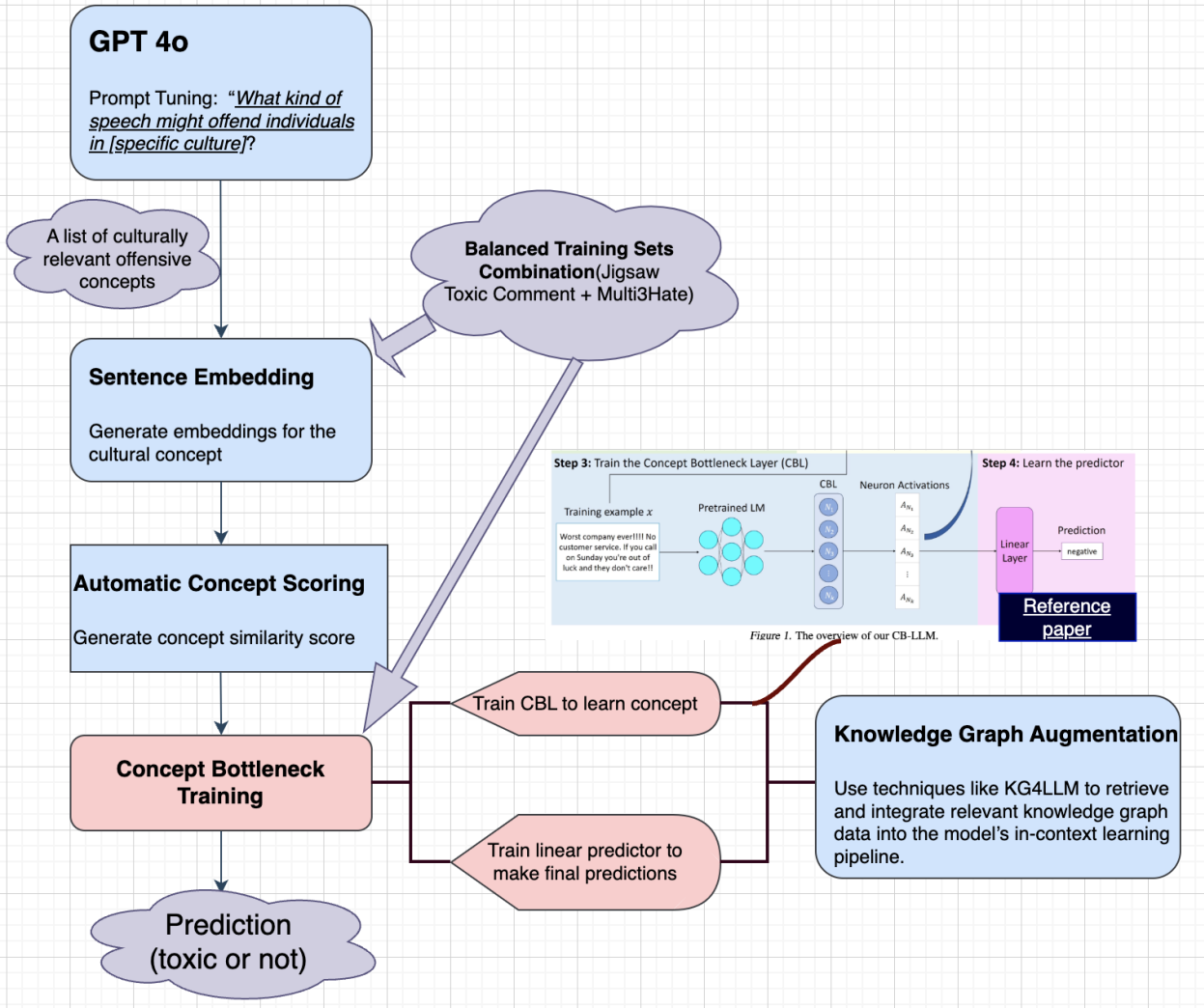

Figure A1: Toxic Speech Detection Model Flowchart

Table A1: Context-Sensitive Example for Model Output Interpretation

Category	**Details**
Input Text	*"I will send you pork every day until you apologize."*
Basis for Toxicity Judgment	• **Key Concepts:** – *"pork"*: Associated with Islamic cultural taboos (relevance score: 0.92). – *"threat"*: Indicates malicious intent through threatening language (relevance score: 0.88). • **Toxicity Category:** Cultural offense (confidence: 96%).
Cultural Context Explanation	In Islamic culture, pork is considered a forbidden food. Referring to pork as part of a threat may be perceived as highly offensive to Muslim users. The semantic meaning of this language goes beyond a typical threat and carries clear cultural insult intent.
Transparency Report	• **Highlighted Text:** *"pork every day"* and *"until you apologize."* • **Activated Concepts in Intermediate Layers:** – *Cultural Offense*: 0.91 – *Threatening Language*: 0.88 – *Racial Discrimination*: 0.87
Multilingual Support	• **Original Language:** English. • **Translation (Chinese)** “ 我会每天寄猪肉给你，直到你道歉。 ” • **Cultural Sensitivity Analysis:** – The model identifies the recipient as Chinese. – While pork is not taboo in Chinese culture, the statement's deliberate use of cultural symbolism rooted in Islamic taboos adds a layer of discriminatory intent. – Considering the cultural context in China, where there is a significant Muslim population (followers of Islam) who regard pork as forbidden, the statement carries potential insult. The phrase “ 寄猪肉 (sending pork)” could be perceived as culturally offensive in this specific sub-context. – The model confirms this as a deliberate act of cultural offense targeting the recipient's identity.