Title

Dynamic Norm Violation Detection: A Cultural Relativism-Inspired Approach for Large Language Models

Problem Statement

Large Language Models (LLMs) often generate responses that unknowingly violate social norms, leading to inappropriate or offensive content. This issue is particularly challenging due to the dynamic and context-dependent nature of social norms across different cultures and contexts.

Motivation

Current approaches to addressing norm violations in LLM outputs rely on static lists of taboo words or topics, or broad toxicity classifiers. These methods lack the nuance and flexibility required to handle the complex, context-dependent nature of social norms. Our approach draws inspiration from anthropological studies on cultural relativism to create a dynamic, context-aware norm violation detector. By leveraging the power of LLMs to learn and adapt to diverse cultural contexts, we aim to develop a more sophisticated and culturally sensitive method for detecting and mitigating norm violations in LLM-generated content.

Proposed Method

We propose a two-stage approach: 1) Norm Learning: Train a model to infer implicit social norms from large-scale social media data across diverse cultures and contexts. This involves developing unsupervised methods to cluster behavioral patterns and identify statistical regularities in social interactions. 2) Violation Detection: Use the learned norm model to evaluate LLM outputs in real-time, flagging potential violations. Implement a few-shot learning mechanism to rapidly adapt to new cultural contexts or evolving norms. The system provides explanations for detected violations and suggests culturally appropriate alternatives.

Step-by-Step Experiment Plan

Step 1: Data Collection

Gather a diverse dataset of social media interactions from multiple platforms (e.g., Twitter, Reddit, Weibo) across different cultures and languages. Ensure the dataset includes a wide range of topics and interaction types.

Step 2: Norm Learning Model Development

Implement an unsupervised learning algorithm (e.g., hierarchical clustering or topic modeling) to identify patterns of behavior and implicit norms in the collected data. Use GPT-4 to generate descriptions of the identified norms for each cluster.

Step 3: Violation Detection Model Development

Fine-tune GPT-4 on the norm descriptions and examples of norm violations. Implement a few-shot learning mechanism to allow rapid adaptation to new contexts.

Step 4: Evaluation Dataset Creation

Curate a test set of LLM-generated responses across various scenarios and cultural contexts. Include both norm-compliant and norm-violating examples. Use human annotators from diverse cultural backgrounds to label the examples.

Step 5: Baseline Implementation

Implement existing methods for comparison: a) Static list-based approach using a predefined list of taboo words and phrases, b) General-purpose toxicity classifier (e.g., Perspective API).

Step 6: Evaluation

Test the proposed system and baselines on the evaluation dataset. Measure performance using metrics such as precision, recall, and F1-score for norm violation detection. Also evaluate the cultural sensitivity and adaptability of each method.

Step 7: User Study

Conduct a user study with participants from diverse cultural backgrounds to assess the system's impact on cross-cultural communication effectiveness. Compare the proposed system against baselines in terms of user satisfaction, perceived appropriateness, and communication outcomes.

Step 8: Error Analysis and Refinement

Analyze cases where the proposed system fails or performs suboptimally. Use these insights to refine the norm learning and violation detection models.

Step 9: Generalization Test

Evaluate the system's performance on a held-out set of cultural contexts not seen during training to assess its generalization capabilities.

Test Case Examples

Baseline Prompt Input

Write a casual greeting to a new colleague in Japan.

Baseline Prompt Expected Output

Hey there! What's up? Nice to meet you!

Proposed Prompt Input

Write a casual greeting to a new colleague in Japan.

Proposed Prompt Expected Output

Norm Violation Detected: The greeting 'Hey there! What's up?' is too informal for a Japanese work context, especially with a new colleague. Suggested Alternative: 'Hajimemashite. Yoroshiku onegaishimasu.' (Nice to meet you. I look forward to working with you.) This greeting is more appropriate in a Japanese professional setting, showing respect while still being relatively casual.

Explanation

The baseline method generates a greeting that, while casual and friendly, is inappropriate in Japanese professional culture. The proposed method detects this norm violation, explains why it's inappropriate, and suggests a culturally suitable alternative.

Fallback Plan

If the proposed method doesn't meet the success criteria, we can pivot to an analysis paper focusing on the challenges of cross-cultural norm detection in LLMs. We would conduct a comprehensive error analysis to identify patterns in the system's failures, potentially uncovering insights about the limitations of current LLMs in understanding cultural nuances. Additionally, we could explore the effectiveness of different unsupervised learning techniques for norm identification, comparing their performance and discussing their strengths and weaknesses. This analysis could provide valuable insights for future research in culturally-aware AI systems and highlight the complexities of encoding cultural knowledge in language models.

Ranking Score: 6