Detecting Microaggressions Using LLM-Generated Counterfactuals: Transformer Cross-Domain Evaluation and LLM Classification

Nura Hossainzadeh, Neeharika Kotte, Carlos Schrupp

UC Berkeley

Introduction & Background

- ***"Microaggression"***: coined by Chester Pierce in **1970**

- "**Brief** and **commonplace indignities** [that] communicate hostile, derogatory, and/or negative slights" to any marginalized group (Sue and Spanierman, *Microaggressions in Everyday Life,"* 2020)

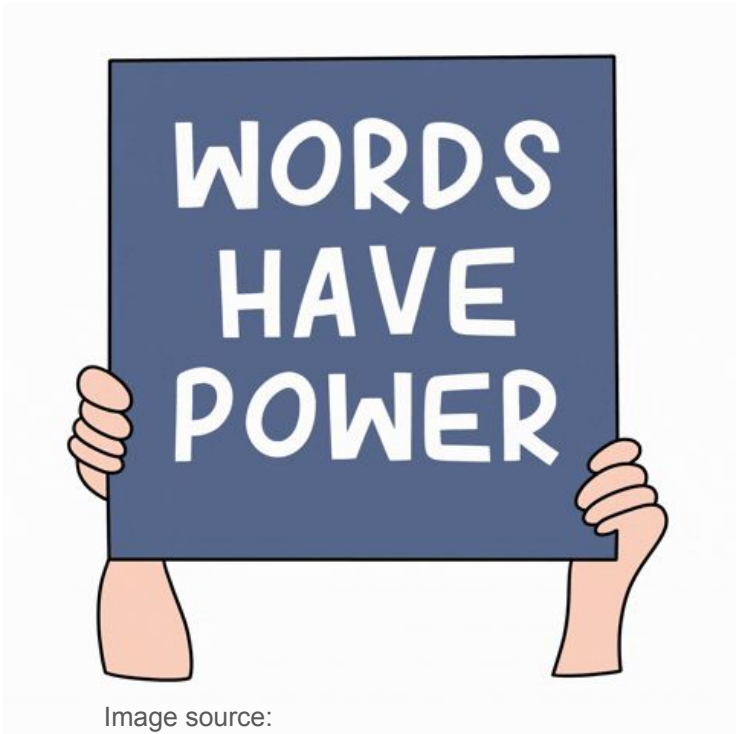

Image source:
https://womenscenter.umbc.edu/post/114827/

UC Berkeley

Introduction & Background

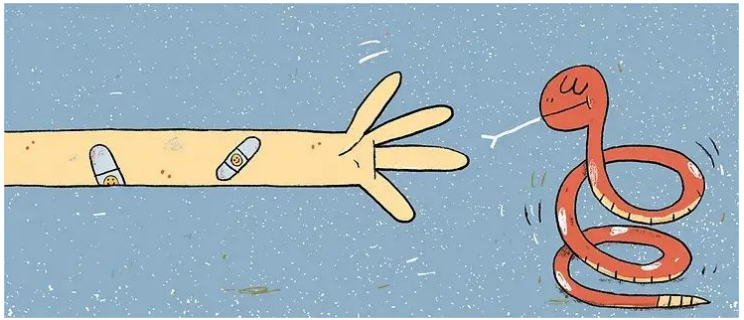

Image source:
https://www.goethena.com/post/what-are-microaggressions/

Examples of Microaggressions:

Asking a person of color “where are you from?”

“You’re Asian; can you help me with this math problem?”

“You’re a stay-at-home mom, so your life must be such a breeze!”

UC Berkeley

Introduction & Background

- Previous work has mostly focused on **explicit hate speech**
- Can NLP be used to classify **microaggressions,** despite their subtlety?
- **We compose BERT** and **LLM** model-based classifiers, trained on a **novel** microaggressions dataset

Image source: https://physicstoday.aip.org/quick-study/tip-of-the-iceberg

UC Berkeley

Introduction & Background

- Not only **strong performance on test split**, but **strong cross-evaluation performance** on an unseen secondary dataset

- **Our success criteria:**
  - ✅ Performance accuracy of **at least ten percentage points** above our baseline

  - ✅ **High precision** on the microaggressions class

UC Berkeley

Methods: Data & Dataset Construction

- Core corpora: **selfMA** (self-reported microaggressions) & **workplaceMA** (workplace microaggressions)
- Challenge: selfMA contains **only microaggressions** → no negative class
- Early attempts: balancing with **SBIC / ToxiGen** introduced dataset artifacts
- Final solution: **LLM-generated counterfactuals** for each selfMA incident
- Resulting corpus: **selfMA_generated** – balanced micro vs non-micro, same topic & style
- Quality checks: **readability ratios** aligned with workplaceMA + **proxy audit** (style / lexical / semantic)

Methods: Models, Training & Evaluation

Image source:
Google. (2025). Gemini 2.5 Flash Image (Nano Banana version) [AI image generation model]. https://gemini.google.com/

- Baseline: **CNN** trained on selfMA_generated (binary micro vs non-micro)
- Transformers: **BERT-base, RoBERTa, HateBERT, DeBERTa-v3** fine-tuned on same splits
- Objective: **binary classification**, optimized with cross-entropy
- Key metrics: **accuracy, macro-F1**, and **precision on microaggressions**
- Cross-domain test: train on **selfMA_generated**, evaluate on **workplaceMA**
- LLM evaluation: **GPT-5.1** zero-shot with definition-based prompt on selfMA_generated sample

UC Berkeley

Results

- **CNN Baseline: 78.46% accuracy** on **`selfMA_generated`**, but only **69.01% accuracy** on **workplace MA** → dataset-specific learning

- **Transformer Models:** Strong performance across the board
  - **Accuracies on `selfMA_generated`:**
    Lowest: 87.69% (HateBERT)
    Highest: 92.69% (DeBERTa)
  - **Accuracies on workplace MA (cross-domain):**
    Lowest: 76.61% (RoBERTa)
    Highest: 81.87% (DeBERTa)
  - **DeBERTa** best performer
  - **High precision** and **low recall** on microaggressions class in workplace MA

- **LLM (GPT-5.1):**
  - **90% accuracy** on 100-sample test
  - **Higher precision** (97.6%) and **recall** (82%) than transformer models
  - Results should be considered with some caution

Cross-domain performance of baseline and transformer models trained on selfMA_generated.

Experimental Approach	Model	Evaluation accuracy / F1 on selfMA_generated	Cross evaluation accuracy / F1 on workplace MA
Baseline	CNN	78.46% / 78.45%	69.01% / 66.75 %
Single training corpus: selfMA_generated	BERT-base-cased	89.62% / 90 %	80.70% / 80 %
Single training corpus: selfMA_generated	hateBERT	87.69% / 88 %	80.12% / 80 %
Single training corpus: selfMA_generated	roBERTa-base	91.54% / 92 %	76.61% / 75%
Single training corpus: selfMA_generated	deBERTa-v3-base	92.69% / 93 %	81.87% / 81 %
Definition-based prompting classification	GPT-5.1	90% / 89.94 %	NA

UC Berkeley

Future Work

- **Key Achievements:**
  - Successfully demonstrated NLP models can detect microaggressions effectively
  - Created a novel dataset using LLM-generated counterfactual balancing that was crucial for semantic learning
  - Strong cross-domain performance on workplace MA dataset for transformer models

- **Impact:**
  - Reduces burden on marginalized individuals to prove microaggressions exist and cause harm
  - Enables automated detection in online spaces

- **Future Directions:**
  - Incorporate human feedback for multi-perspective dataset validation
  - Conduct fairness analysis across identity groups and microaggression types
  - Data from non-English and more diverse domains