Evaluating the Impact of Explicit Hate Speech Definitions on the Stability of LLM-based Hate Speech Classification

Rodrigo Santos - R.M.MartinsdosSantos@student.tudelft.nl

Supervisors: Pradeep Murukannaiah, Urja Khurana

EEMCS, Delft University of Technology, The Netherlands

1) Introduction

- Evaluating Hate Speech using LLMs can lead to inconsistent evaluations, despite maintaining similar accuracy.

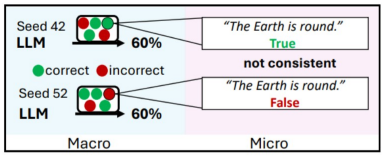


Figure 1: Macro-consistency vs Micro-consistency

- Aleatoric uncertainty: Inherent randomness of the data
- Epistemic uncertainty: Model's limited information

Figure 2: Aleatoric vs epistemic uncertainty

- Modern neural networks are poorly calibrated, making them overconfident

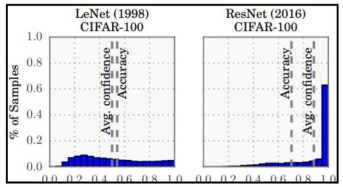


Figure 3: Overconfidence of modern Neural Networks

2) Research Question

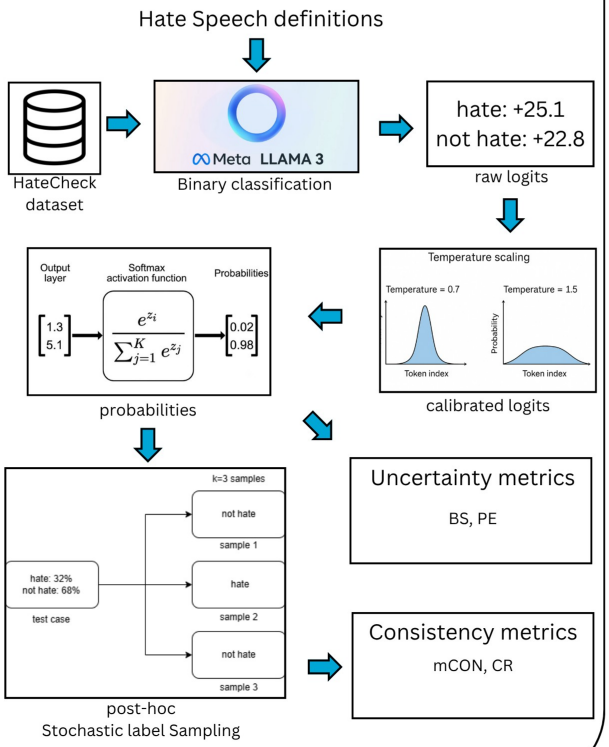
"Does providing a definition of hate speech to an LLM improve the micro-consistency and uncertainty on hate speech classification?"

3) Methodology

Metrics

- **Predictive Entropy (PE)**: total uncertainty of model predictions
- **Brier Score (BS)**: total uncertainty and calibration.
- **Consensus Ratio (CR)**: Calculates micro-consistency across samples using the majority vote.
- **Mean Consistency (mCON)**: Measures micro-consistency across samples using pairwise comparisons.

$$mCON(S) = \frac{1}{\binom{|S|}{2}} \sum_{i<j} CON(S_i, S_j)$$



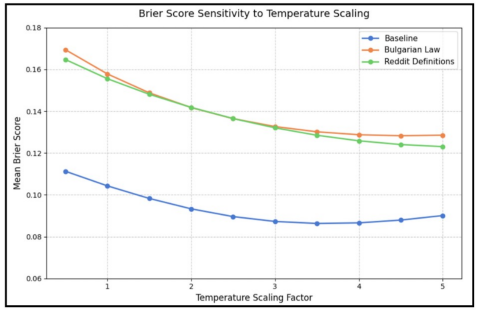
4) Results

Hate Speech Definition	Optimal Temperature (T)
Baseline	3.145
Bulgarian Law	4.345
Reddit Definition	5.012

optimal temperatures from minimizing NLL

Metric	Baseline	Bulgarian law	Reddit
Brier ↓	0.104	0.158*	0.156*
PE ↓	0.591	0.596*	0.592*
CR ↑	0.971	0.954*	0.967*
mCON ↑	0.958	0.934*	0.952*

metric mean per definition at T=1



Brier Score relation to temperature (k=50)

Definition	Metric Pair	Correlation (r)
Baseline	BS vs. PE	0.96
	BS vs. mCON/CR	-0.58
	PE vs. mCON/CR	-0.67
Bulgarian	BS vs. PE	0.86
	BS vs. mCON/CR	-0.60
	PE vs. mCON/CR	-0.77
Reddit	BS vs. PE	0.89
	BS vs. mCON/CR	-0.52
	PE vs. mCON/CR	-0.68

Spearman's rank correlations

5) Conclusion and Future Work

- The results suggest that the LLM was unable to properly adapt to explicit definitions of hate speech.
- Micro-consistency and uncertainty are strongly correlated.
- Further research is required to prove that the task is a fundamental challenge for current models, rather than a limitation of specific models.

Future Work

- Apply MC dropout to isolate epistemic uncertainty.
- Try alternative configurations (such as LoRA) to improve pre-calibration logits
- Run experiments with more explicit definitions