UNIVERSIDADE DA CORUÑA

Llama-3-8B-Distil-MetaHate

Usage/License

The model can be used via the Hugging Face transformers library. It was released under the llama3 license.

Download

The model can be accessed and downloaded from its Hugging Face page:

https://huggingface.co/HYBRIDS/Llama-3-8B-Distil-MetaHate

Researchers

- Paloma Piot
- Javier Parapar

These researchers are affiliated with the Information Retrieval Lab at the University of A Coruña.

Summary

Llama-3-8B-Distil-MetaHate is a distilled version of the Llama 3 architecture, fine-tuned for hate speech detection and explanation. Developed by the Information Retrieval Lab at the University of A Coruña, this model employs Chain-of-Thought reasoning to enhance interpretability in hate speech classification tasks. The model aims to not only detect hate speech but also provide explanations for its classifications.

Key Features

- Distilled Llama 3 Architecture: A more efficient version of the original Llama 3 model, fine-tuned for hate speech classification and explanation.
- Chain-of-Thought Reasoning: Provides insights into the decision-making process by explaining why a piece of text is classified as hate speech.
- Improved Efficiency: Uses model distillation to retain high accuracy while reducing computational costs.
- Binary Classification & Explanation: Outputs both a classification label (hate/no hate) and an accompanying textual explanation.

What We Offer

This model delivers a more efficient and explainable approach to hate speech detection by leveraging model distillation techniques. It is particularly useful for researchers, content moderation platforms, and policymakers looking for a transparent and interpretable solution in online harm detection.

Collaboration Objectives

This model aligns with broader exploitation strategies to enhance explainability in hate speech detection models, reduce computational costs while maintaining accuracy, foster collaborations with research groups and industry partners interested in improving hate speech moderation techniques and investigate bias mitigation strategies in hate speech classification.

hybrids

HYBRIDS Website

www.hybridsproject.eu

Get in Touch

paloma.piot@udc.es