MetaHate BERT

Usage/License

The model can be used via the Hugging Face transformers library. The model is licensed under the Apache 2.0 License. Users are advised to be aware of potential biases present in the training data and understand that the model may misclassify some instances. Performance may vary across different domains.

Download

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

https://huggingface.co/HYBRIDS/MetaHateBERT

Researchers

- Paloma Piot
- Patricia Martín Rodilla
- Javier Parapar

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

Summary

MetaHateBERT is a fine-tuned BERT model specifically designed to detect hate speech in text. It is based on the bert-base-uncased architecture and has been trained for binary text classification, distinguishing between 'no hate' and 'hate'.

Key Features

- Hate Speech Detection: Accurately classifies text as "hate" or "no hate".
- Content Moderation: This model can be used to assist platforms in automatically flagging potentially harmful content.

What We Offer

A pre-trained model capable of identifying hate speech in various textual data sources, such as social media comments and forum posts.

Collaboration Objectives

As part of the broader MetaHate Collection and Exploitation Strategy, the goal is to foster collaboration with academia and industry. The objectives include:

- Enhancing the model’s performance across diverse domains and languages.
- Sharing expertise to refine hate speech detection methodologies.
- Building a unified framework for tackling hate speech using state-of-the-art NLP techniques

hybrids

HYBRIDS Website

www.hybridsproject.eu

Get in Touch

paloma.piot@udc.es