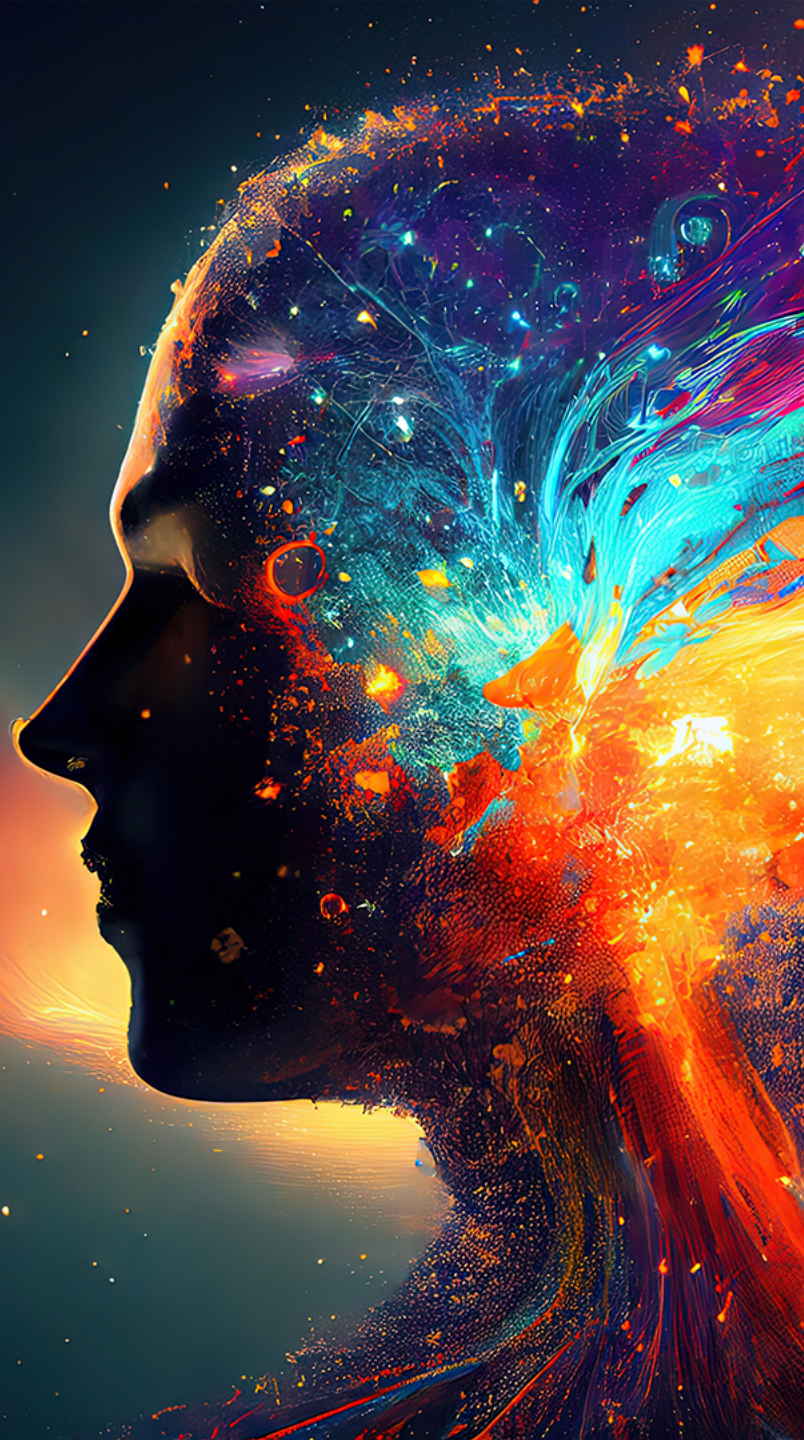
Analytics Vidhya

Chain-of-Explanation (For Explaining Hate Speech)



Instructor

Bhaskarjit Sarmah

Vice President, Blackrock

Definition

- Chain-of-Explanation method offers detailed insights into potential hate speech by highlighting not-so-obvious cues.

- This approach begins by identifying trigger words and the targeted group in subtly harmful texts, aiming to generate clear, educative explanations.

Analytics Vidhya

How Chain-of-Explanation Works

A sequential methodology guides the language model to produce precise explanations from initial, nuanced clues.

Trained with diverse examples, the model adopts this structured sequence to create explanations that are informative yet straightforward to comprehend.

The quality of the explanations is gauged against human judgement, emphasizing the clarity and informativeness of the output.

Analytics Vidhya

Thank You

Analytics Vidhya