AI FACTSHEET

Model

Name: SentimAI

Author Notes

Ensemble: SentimAI incorporates an ensemble of transformer-based models to enhance sentiment detection accuracy across languages.
Robustness: Extensive testing for robustness against common adversarial attacks like FGSM and DeepFool was performed, with adjustments made to improve resilience.

Overview

Document Summary: This FactSheet accompanies the SentimAI model, designed for comprehensive sentiment analysis in multiple languages for marketing and customer feedback analysis.
Purpose: To classify the sentiment of given text across multiple languages, aiding in understanding customer feedback and social media analysis.
Intended Domain: Natural Language Processing, with specific application in marketing analysis and customer feedback interpretation.

Training Data

Dataset Used: Utilises the Multi-Domain Sentiment Dataset combined with a curated collection of customer feedback from various online platforms. The total dataset encompasses over 150 million sentences.
Preprocessing: Includes normalisation, tokenization, and removal of sensitive information using automated scripts.

Model Information

Architecture Description: SentimAI uses a two-stage transformer-based architecture. The first stage generates embeddings from text input, and the second stage classifies sentiment based on these embeddings.
Input Output Process: The model accepts raw text input and outputs sentiment scores ranging from -1 (negative) to 1 (positive) alongside confidence levels.

Inputs and Outputs

Inputs: Raw text in English, Spanish, French, or German, with a recommended maximum of 512 tokens.
Outputs: A JSON object containing sentiment scores and associated confidence levels for the input text.

Performance Metrics

Metrics Used: Accuracy, Precision, Recall, F1 Score, and Mean Average Precision (MAP).
Results: SentimAI achieved an average accuracy of 92% across all supported languages, with an MAP of 0.357 and an Area Under the Curve (AUC) of 0.968.

Bias

Potential Biases: An ongoing review process has been established to identify and mitigate potential biases, especially concerning the representation of dialects and informal language.

Robustness Tests

Attack Resilience: Demonstrated resilience against L-infinity and L2 norm attacks, with detailed performance metrics available upon request.

Domain Shift

Evaluation: Continuous monitoring and evaluation mechanisms are in place to assess the model's performance against shifting data distributions, ensuring sustained accuracy and reliability.

Test Data

Description: The test set includes unseen text from various online forums and review sites, maintaining a diverse representation of sentiment and language use.
Split Ratio: Data was split into 70% training, 20% validation, and 10% testing segments.
Class Ratio Maintenance: Efforts were made to maintain consistent class ratios across all data splits to ensure model fairness.

Operational Conditions

Optimal Conditions: The model shows optimal performance in high-quality text inputs containing clear sentiment expressions.
Poor Conditions: Performance decreases in texts with mixed sentiments, slang, or significant noise.

Explanation

Model Explainability: Despite the inherent complexities of deep neural networks, SentimAI employs layer-wise relevance propagation (LRP) techniques to provide insights into decision-making processes.

Contact

Information: For more information or queries regarding SentimAI, please reach out to the development team at info@sentimai.com.