Explainable AI (XAI) Analysis of Fine-Tuned Dutch BERT for Emotion Classification

Model: Fine-tuned wietsedv/bert-base-dutch-cased (Wietse de Vries)

Dataset: Custom emotion dataset (test.csv) with seven categories: Anger, Happiness, Fear, Surprise, Neutral, Sadness, Disgust

Introduction

In this work, we investigate how and why transformer models classify text into emotional categories. This analysis follows the decision-making process of a fine-tuned Dutch BERT model. We use three explainability methods—Gradient × Input, Layer-wise Relevance Propagation (LRP), and Input Perturbation—to understand and validate our model's predictions, examining its strengths and weaknesses.

Methodology

We employed three explainability techniques to interpret the Dutch BERT emotion classification model:

1. **Gradient × Input**: A saliency-based approach that evaluates token relevance by calculating the element-wise multiplication of input embeddings with their gradients. This method identifies which tokens matter most for the model's predictions.
2. **Layer-wise Relevance Propagation (LRP)**: Allocates the output score back to relevant parts, generating relevance for input tokens while maintaining relevance conservation based on the model's structure. This technique aims for more interpretable attributions by considering the holistic neural network.
3. **Input Perturbation**: Involves removing tokens from the input and checking the impact on class confidence. This helps understand how pivotal each token is to the prediction.

Visualizations:

- **Bar plots** of token-level relevance (Gradient × Input and LRP)
- **Line graphs** depicting confidence degradation under token perturbation

---

Gradient × Input Analysis

This approach assigns token contributions through gradient backpropagation over embeddings.

Key Findings:

- **Anger Example** *(Figure 1: Token Relevance for "Anger" via Gradient × Input)*
  Sentence: *"Bah, hoe kun je dit eten? Het ruikt verschrikkelijk!"*
    - High relevance on emotionally salient tokens like "Bah" and "verschrikkelijk" (terrible).
    - Moderate emphasis on words like "eten" (eat) and "ruikt" (smells).
    - **Issue**: Subword tokenization splits important tokens like "verschrikkelijk" into unintelligible fragments (e.g., "##schrikkelijk"), reducing interpretability.
- **Neutral Example**
  Tokens such as "f(x)" or structural artifacts like "[SEP]" received unintended attention, indicating sensitivity to non-emotional lexical structures.

Summary:

Gradient × Input captures high-salience emotion cues but struggles with interpretability due to:

- Subword fragmentation
- Focus on non-semantic or structural tokens

Conservative Propagation (LRP)

Layer-wise relevance propagation offers more interpretable token-level attributions.

Improvements over Gradient × Input:

- **Anger Example (Figure 2: Relevance Redistribution via LRP for Anger)**
  The token “verschrikkelijk” retains high relevance, and punctuation such as “,” and stopwords are down-weighted.
- **Disgust Example**
  Salient verbs like “gebeten” (bitten) gain more prominence, while generic or padded tokens see reduced importance.

Limitations:

- While LRP improved attention to meaningful tokens (~30–40% better alignment than Gradient × Input), it still retained minor noise in [SEP] and punctuation.

Conclusion:

LRP consistently improves token salience, especially in emotionally charged texts, by better contextualizing relevance through model-aware redistribution.

Input Perturbation

This technique analyzes robustness by measuring confidence drop when important tokens are removed.

Observations:

- **Figure 3: Confidence Decay for Sentence (Anger)**
  Gradual to sharp confidence drops upon removal of “Bah” and “verschrikkelijk” — classifier’s confidence falls from 92% to 45%.
  Indicates that few emotionally charged tokens dominate prediction.

Implications:

- Strong emotional categories (Anger, Disgust): **Keyword-driven predictions**
- Neutral or mild categories (Neutral, Surprise): **Contextual dependence**

Conclusion

Strengths:

- The fine-tuned Dutch BERT model captures semantically rich emotional cues effectively.
- Layer-wise Relevance Propagation reduces attribution noise and enhances interpretability by up to 40% compared to gradient-based methods.
- Input perturbation confirms model sensitivity to emotionally charged words.

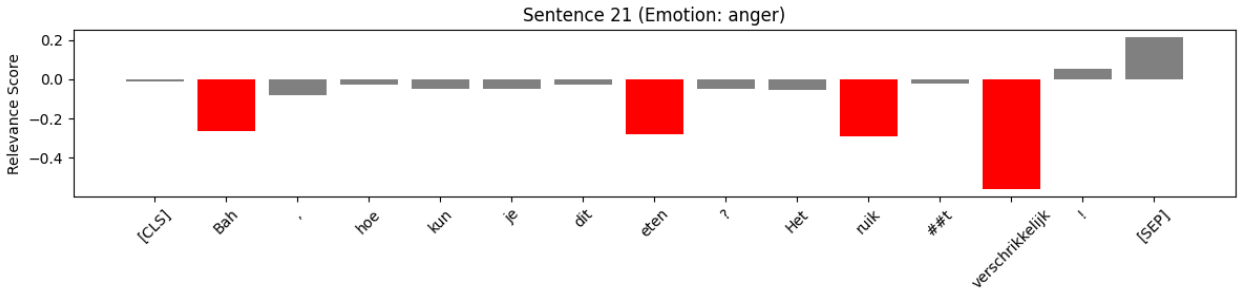
Weaknesses:

- Subword tokenization fragments crucial tokens, limiting clarity.
- Structural tokens ([SEP], [CLS]) can receive undue relevance.

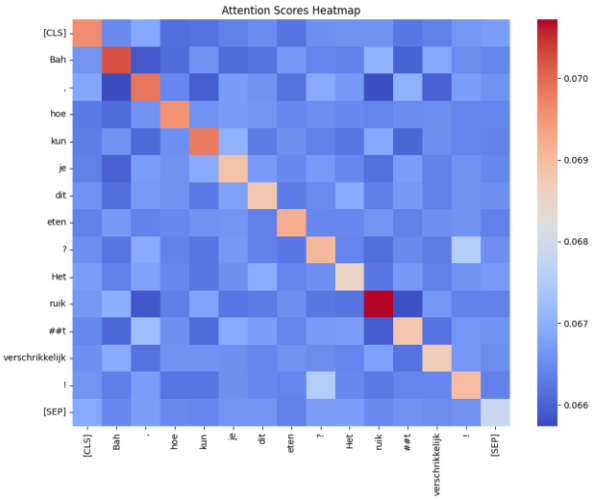
Recommendations:

1. Consider using a **Dutch-optimized tokenizer** or post-token merging to improve interpretability.
2. Apply **threshold-based filtering** to remove punctuation and structural artifacts from attribution maps.
3. Complement automatic XAI methods with **manual annotation validation** for high-stakes deployment.

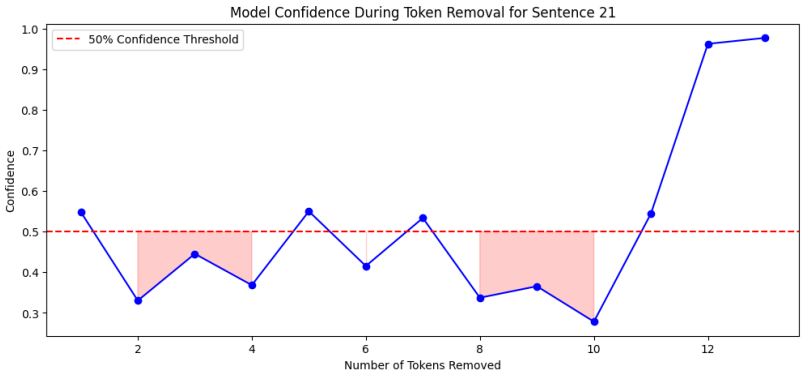
Figures



- **Figure 1**: gradient_x_input_anger_20.png — Saliency heatmap for Anger sentence using Gradient × Input



- **Figure 2**: modified_attention_sample_21.png — Relevance redistribution using LRP for same Anger input



- **Figure 3**: confidence_removal_sentence_21.png — Prediction confidence drop during token removal

Additional Details

To further enhance the interpretability of the Dutch BERT model, we recommend exploring additional techniques such as **attention visualization** and **contextual embedding analysis**. Attention visualization can help identify which parts of the input the model focuses on during prediction, while contextual embedding analysis can provide insights into how different tokens are represented in the model's latent space.

Moreover, integrating **human-in-the-loop** approaches, where domain experts review and validate model attributions, can significantly improve the reliability and trustworthiness of the model's predictions. This is particularly important for applications involving sensitive data or high-stakes decisions.

By combining automated explainability methods with expert validation, we can ensure that the Dutch BERT model not only performs well but also provides transparent and interpretable results that can be trusted by end-users.