Model Card – Emotion Classification for Unscripted TV Program

Model Details:

- Developed by Data Science and AI students at Breda University of Applied Sciences, Team 28 Y2C
- Fine-tuned DistilRoBERTa-Base (Liu et al., 2019)
- Model type: Sequence classification transformer-based model
- Submitted on huggingface.co on Oct 17, 2019
- Paper for the base model: https://arxiv.org/abs/1907.11692
- Huggingface page for the base model
- License: Apache 2.0

Intended Use:

- Meant to be used for tasks of emotion classification of 6 core emotions, such as happiness, anger, surprise, disgust, sadness and fear, + neutral
- Developed for the analysts of tv shows with unscripted replicas but can be used for similar tasks of providing emotion per sentence.
- Model was trained on both spoken and written text, so using it for classification of both sequence origins should not bring any issues.
- This model is fine-tuned for sequence classification. Model expects sentences under 512 tokens, as an input from tokenizer. For tasks such as text generation you should look at a model like GPT

Factors:

- Variations in how people express emotions across different languages, dialects, or regional slang could impact performance. DistilRoBERTa is trained primarily on standard English, so informal, coded, or non-standard expressions may be misclassified.
- Multi-meaning or multi-label sentences will have only one emotion, as the model is not meant to grasp all emotions within the context.
- Some emotions (e.g., happiness, neutral) might have more training examples than others (e.g., disgust, fear). This leads to skewed predictions where underrepresented emotions are harder to detect.

Training and Evaluation Data:

	GoEmotions	**Other teams' data**	**Own data**
Used for	Train/Val	Train/Val	Test

Samples	54,257	31,967	1042
Distribution of row length (in characters)	mean: 65.5 std: 35.67 range: 2 - 397	mean: 30.7 std: 21.67 range: 1 - 516	mean: 35.1 std: 21.15 range: 1 - 151
Happiness	20,345	385	6,688
Neutral	16,885	398	12,425
Anger	6,182	37	3,071
Surprise	5,766	186	7,479
Sadness	3,441	20	1,300
Disgust	835	12	839
Fear	809	5	165

- GoEmotions used solely for training. *Dataset of written comments manually labelled with emotions, which best expected data to train in relation to the intended use, but still helpful to grasp the context*.
- Test dataset provided by the company. This dataset is made up of transcribed spoken text, which is the intended dataset to test on.
- Additional training data collected from other teams (their test sets of spoken data) to mix with the written data to make the model uniform for both scenarios.

Metrics:

The goal was to see how the model performs within this imbalanced dataset, that is why f1-macro and f1-weighted were used to see the performance of under-represented classes and in general.

Emotion	Precision	Recall	F1-Score	N of samples
Neutral	0.66	0.79	0.72	399
Happiness	0.8	0.72	0.76	385
Surprise	0.73	0.62	0.57	186
Anger	0.51	0.51	0.51	37
Sadness	0.55	0.55	0.55	20
Disgust	1.00	0.17	0.29	12
Fear	0.5	0.40	0.44	5

The model achieves **0.71 f1-weighted** with a **macro F1-score of 0.56 on the test set**, highlighting performance imbalances due to class distribution. While **happiness (F1: 0.76)** and **neutral (F1: 0.72)** perform well due to higher support, **disgust (F1: 0.29, recall: 0.17)** and **fear (F1: 0.44, recall: 0.40)** suffer from low recall, indicating underrepresentation. The **confusion matrix** shows **neutral is often over-predicted**, likely due to cautious classification, while **surprise is misclassified as neutral or happiness**, suggesting overlap in textual features.

Variability during training: macro f1: 0.5760 ± 0.0086, weighted f1: 0.6595 ± 0.0019, measured using StratifiedKFold technique with 5 splits to account for irregularities in training/validation splits.

Error Analysis:

The model's errors exhibit **structured patterns**, primarily involving misclassifications between **happiness, neutral, and surprise**, which dominate the dataset. These confusions are often caused by **lexical overlap and emotional ambiguity**, especially in short or generic sentences. Despite leveraging **self-attention mechanisms** inherent to the Transformer architecture, the model still struggles with **context-poor inputs**, where emotional intent is subtle or implied. Rare emotions like **disgust, fear, and anger** are often under-predicted due to severe **class imbalance** and the lack of explicit emotional markers in text. Short inputs often lack enough semantic cues, while longer ones tend to contain **emotionally coloured words** that mislead the model despite a neutral or factual tone. This suggests that while attention captures intra-sentence relationships, the model may benefit from **extended context** across sentence boundaries (e.g., conversational history) and **better lexical disambiguation**. Improvements could involve **class rebalancing**, **training on richer linguistic variations**, and **augmenting data** with more expressive examples for underrepresented emotions.

XAI:

The current model's XAI concerns centre on **limited transparency** and **fragile interpretability**, particularly for underrepresented emotions. While LRP methods reveal reliance on key tokens (e.g., "gorgeous" for happiness, "complaints" for anger), **class imbalance** severely impacts explanations for rare classes like disgust (F1: 0.29) and fear (F1: 0.44), where low recall and sparse training data lead to weak or inconsistent token relevance signals. Surprise predictions suffer from **distributed token reliance**, making explanations ambiguous, while neutral over-prediction suggests **contextual blindness** in emotionally subtle inputs. Furthermore, **biases from the base model's training on unfiltered English data** risk misclassifying non-standard expressions, yet XAI methods fail to highlight these systemic gaps. The discrepancy between Gradient × Input (noisy) and LRP (clearer but incomplete) underscores **methodological limitations**, potentially misleading users about model robustness. These issues threaten trust in critical applications, as explanations may not reliably reflect the model's decision logic across diverse or nuanced scenarios.

Ethical Considerations:

As it is mentioned in Huggingface description of the model, it is trained on Wikipedia and internet non-filtered data, which can be skewed and biased. For example, they tried to predict next word (profession) for people of different races and the answers were biased. Regarding human-related risks, it is not recommended to use this model, as a main source of analysis for the fields like mental health assessment. In order to manage such issues, we trained the model on task specific dataset for better contextual representation and suggest using it in intended way.

Caveats and Considerations:

- Suggested to do a more thorough testing on underrepresented classes
- All tasks containing emotions will always have some part of error, as emotion are interpreted differently from sentence tom sentence and from one person to another, which brings some inconsistencies into the data.
- Requires additional fine tuning for more domain specific utilization within sensitive applications.