1. Introduction

Emotions play a crucial role in human interaction and communication. They influence the meaning, intonation and interpretation of what is said. However, for Natural Language Processing (NLP) models, the task of automatically detecting emotion from text remains extremely challenging. This is especially true in the context of spoken language, where emotional colouration is often expressed through intonation, pauses, facial expressions or context rather than clear words.

As part of an emotion categorisation project, we have developed and tested a model trained on transcriptions of TV shows. This data represents spontaneous speech translated into text, making the task even more challenging. The model had to determine which of the following seven emotions was expressed in each sentence: **anger, disgust, fear, happiness, sadness, surprise, neutral**.

One of the best models, **DistilRoBERTa**, further trained on our marked sample, was used for the analysis. The main purpose of this report is to analyse the errors of the model on a test dataset and identify key patterns that can be used for its further improvement. We consider both quantitative aspects (accuracy, completeness, F1-metrics) and qualitative aspects (sentence structure, length, emotional ambiguity, etc.).

2. Overall performance of the model

On the test set of 1044 rows, the model performed quite confidently, especially in real-world applications:

- Accuracy (total proportion of correct predictions): 71%
- Macro F1-score (average F1 over all classes): 0.56
- Weighted F1-score (weighted across classes): 0.71

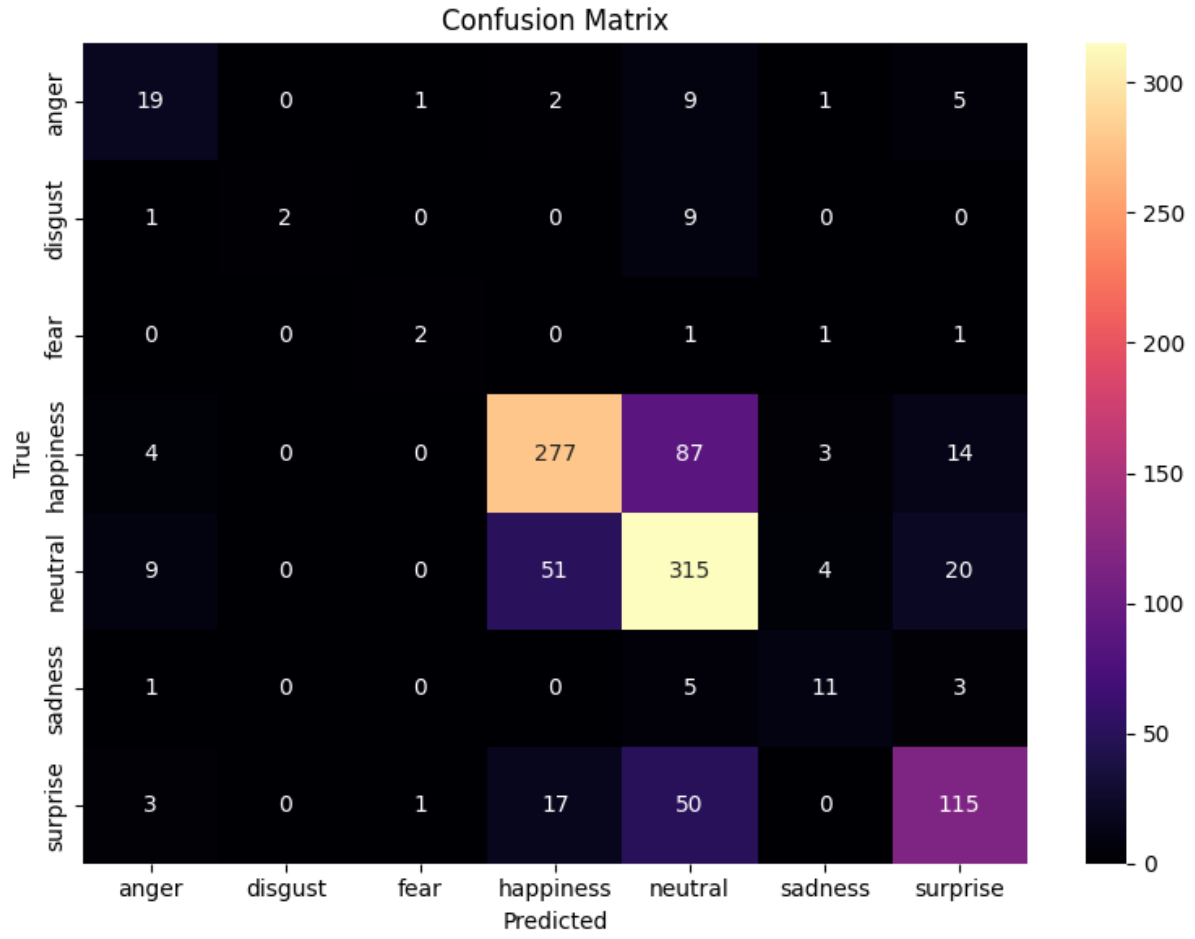
This means that despite certain difficulties, the model is already able to recognise emotions in real spontaneous speech quite well. The frequently occurring classes **neutral** and **happiness** are particularly well recognised.

However, the breakdown by class shows a strong imbalance in performance:

- Class **disgust**: F1-score of only 0.29 with an accuracy of 1.00, but recall of 0.17
- Class **fear**: F1-score 0.44.
- While **happiness** achieves an F1-score of 0.76 and **neutral** achieves an F1-score of 0.72

Thus, the average metrics mask poor performance on less represented emotions, and this is one of the main sources of model error.

3. Analysing the error matrix (confusion matrix)



The error matrix of the model showed the presence of consistent, recurring patterns of misclassification. The most frequent cases of errors are given below:

True label	Predicted label	Number of errors
happiness	neutral	87
neutral	happiness	51
surprise	neutral	50
neutral	surprise	20
happiness	surprise	14

Such errors are indicative of several key problems:

1. Emotional bias - the model tends to overestimate **neutral**, erroneously reducing positive or surprised emotions to it
2. Class similarity problem - **happiness**, **neutral**, **surprise** often have similar lexemes, which biases the model.
3. The model 'plays it safe' - especially on short sentences it often predicts **neutral** by default.

Separately, it is worth noting that most errors occur not between diametrically opposed emotions (e.g. **happiness** ↔ **anger**), but between similar in tone: **neutral** ↔ **surprise**, **happiness** ↔ **neutral**.

4. Frequent errors and patterns

We analysed the top 10 most frequent combinations of errors (true_label → predicted_label). Interestingly, almost all of them occur between the three main classes: **happiness**, **neutral**, **surprise**.

This may be due to several factors:

- The frequency of these emotions in the dataset
- The absence of explicit emotional markers
- The prevalence of short, mundane lines in the dataset

It is also important to note that:

- **anger**, **disgust** and **fear** are rarely predicted by the model - even when they are set as ground truth.
- The **surprise** → **happiness**, **happiness** → **surprise** errors often occur on longer and richer descriptions, especially in memory stories where emotions are mixed.

These patterns make us realise that model errors are not random but have a logical structure. This allows us to take a point-by-point approach to improving it, rather than just 'adding more data.'

5. Effect of sentence length on model errors

One of the hypothesised factors affecting model errors is sentence length. To confirm or refute this, we further analysed all classification errors by dividing the sentences into three groups:

- Short (up to 5 words)
- Medium (6-10 words)
- Long (11 words or more)

Error distribution:

- Short: 100 errors
- Medium: 136 errors
- Long: 67 errors.

This means that almost half of all errors (about 47%) occur on sentences up to 10 words long, i.e. short and medium. In turn, long sentences (more than 10 words) account for only a third of all errors. However, in long sentences, errors are more often related to ambiguity of emotion and confusion of meaning.

❖ Examples of short sentences with errors:

1. **'restaurant singer anastasia ponomareva'**

→ True: **neutral** | Predicted: **happiness**

(a one-word topic is often interpreted as positive)

2. **'in addition i accept guests'**

→ True: **neutral** | Predicted: **happiness**

(the word ‘guests’ can evoke a positive association)

3. **‘mom always baked pies.’**

→ True: **happiness** | Predicted: **neutral**

(without context, the emotional connotation is lost)

❖ Examples of long sentences with errors:

1. **‘the task of everyone is to demonstrate to guests all their talents’**

→ True: **neutral** | Predicted: **happiness**

(words like ‘talents’ and ‘guests’ conjure up happiness)

2. **‘the winner of the week will become a participant in the superfinal.’**

→ True: **neutral** | Predicted: **happiness**

(the model focuses on the positive colouring of words, ignoring the absence of emotion)

3. **‘well since I still have building a house for my mother’**

→ True: **happiness** | Predicted: **neutral**.

(long structure detracts from the emotional focus)

Conclusion: the model does not have a stable advantage on either short or long sentences. Errors on short sentences are often due to lack of context, while on long sentences they are due to meaning overload. Particularly frequent are cases where the model interprets neutral or descriptive phrases as **happiness**, indicating a potential bias towards positive interpretation.

6. Strengths and weaknesses of the model

Strengths:

- High accuracy on happiness and neutral classes, which are the most frequent.
- The model performs well on positive emotions and calm-descriptive phrases.
- Relatively confidently classifies surprise despite its semantic fuzziness.

Weaknesses:

- The classes disgust, fear, anger are almost unpredictable: the model favours ‘safe’ emotions.
- Errors between neutral ↔ happiness and neutral ↔ surprise are the most frequent.
- The model does poorly with sarcasm, circumlocution, and indirect forms of expressing emotion.
- The high recall of neutral indicates its overuse: even emotional phrases are often mislabelled as neutral.

- The absence of lexical or syntactic signs of emotion leads to systematic errors (e.g. phrases mentioning guests, food, victory are treated as happiness by default).

The problem is also that the model relies solely on lexical composition without access to intonation, facial expressions or dialogue context, which limits its ability to accurately interpret emotions.

7. Which is more important - Accuracy, Recall or F1-score?

Although the overall accuracy of the model is 71%, it does not give a complete picture of the real performance. Why? Because there is a strong class imbalance in the emotion classification task: some emotions (e.g. **neutral**, **happiness**) are 5-10 times more common than **disgust** or **fear**.

If the model always predicts **neutral**, it will already achieve high accuracy. However, its ability to discriminate between less frequent but meaningful emotions will be nullified.

In this situation, the following are particularly important:

- Recall (Completeness): how well the model finds all instances of a particular emotion. Critical for **disgust**, **fear**, **anger**, where errors can have a meaningful effect - especially in media content where these emotions determine tone.
- Macro F1-score: the average F1 value across all classes, without considering their frequency. This is the most honest metric in an imbalanced environment because it equates the importance of each emotion.

Thus, for our task, it is Macro F1-score that is the key metric, and the current value of 0.56 indicates the need to improve the model on weak classes.

8. Linguistic features of the errors

The error analysis showed that the model is often wrong in cases where:

- Emotion is expressed indirectly (circumlocution, metaphor)
- Restrained wording is used (especially in fear, anger, disgust)
- Positive language is misleading, even if there is no emotional connotation

Examples:

- **‘the winner of the week will become a participant in the superfinal’**

→ True: **neutral** | Predicted: **happiness**

Reason: positively coloured words mislead the model, although the emotion is neutral (fact)

- **‘mom always baked pies.’**

→ True: **happiness** | Predicted: **neutral**

Reason: without the context of childhood and warmth, the model fails to capture the emotion

- **‘they eat her hands’**

→ True: **neutral** | Predicted: **anger**

Reason: strange construction can be perceived as aggression

Frequent errors were also recorded in sentences with common words: **guests, dinner, smile, cheese, win, family**. They may occur in different contexts, but the model often automatically associates them with **happiness**.

Conclusion: the model relies excessively on lexical patterns and is unable to distinguish the context of word usage. This is critical, especially in emotion categorisation, where the same words can convey opposite states depending on intonation or situation.

9. Recommendations for Content Intelligence Agency

Based on the error analysis, we developed the following recommendations to improve the model:

Architectural and technical:

- Use audio fiches (pitch, intonation, pauses) in the pipeline, especially for the emotions **surprise, fear, disgust**, which are often expressed non-verbally
- Refine the model on additionally labelled rare class data
- Add an attention mechanism to the preceding and following sentence to take context into account

Methods for improving data quality:

- Manual filtering and balancing of classes in the training sample
- Data augmentation with reformulations: different formulations of the same emotion
- Use of external emotion datasets, including sets labelled for sarcasm/irony

For the client:

- Clearly mark-up scenarios where errors are most critical (e.g., analyses of conflict scenes, disturbing themes)
- Use the model as an auxiliary tool, not as the sole source of markup

10. Conclusion

This analysis demonstrated that our **DistilRoBERTa**-based emotion classification model achieves good results in terms of accuracy metrics and adequate performance on frequent emotions (neutral, happiness). However, its ability to recognise less frequent but meaningful emotions (such as fear, disgust, anger) requires improvement.

The errors of the model are systematic:

- Frequent substitutions between **neutral, happiness, surprise**
- Strong influence of positive vocabulary on predictions
- Inattention to sentence structure and context

Nevertheless, such patterns mean that the model can be improved point by point - with balanced data, additional features (including audio) and training strategies.

Implementing the recommendations suggested in this report will help the Content Intelligence Agency create a more robust and interpretable tool for analysing emotion in media content.