Technical Report – Emotion Classification Pipeline

Prepared for:

Banijay Benelux

Prepared by:

Andrea Tosheva, Dominik Szewczyk, Rebecca Borski, and Borislav Nachev

Breda University of Applied Sciences

12.04.2024

Technical Report – Emotion Classification Pipeline

Abstract

This abstract describes the development of an emotion classification pipeline for the TV series Expeditie Robinson using advanced Natural Language Processing (NLP) techniques, particularly leveraging BERT models. The pipeline achieved an overall accuracy of 53%, with strong performance in identifying happiness but challenges in recognizing other emotions like disgust and surprise. Error analysis highlighted the influence of word types and text features on predictions, suggesting areas for improvement.

Key challenges included dependence on translated text and limited domain-specific training data. To enhance the pipeline's performance, future recommendations include gathering more specific and longer data tailored to Expeditie Robinson, training with Dutch language datasets using models like RobBERT, and refining the labelling process for improved accuracy.

Keywords: AI, NLP, BUas, Expeditie Robinson, Bert, RoBERT, emotion classification

1. Introduction

Emotions can be represented by language, vocal intonation, nonverbal clues, or written content. They are subjective experiences that reflect an individual's affective state. In recent years, machine learning techniques have been adopted to identify and categorize emotions. Emotion detection finds several practical applications in fields like marketing, where opinion mining evaluates customer sentiment toward products or services. With automated emotion analysis, managers can quickly and effectively extract emotional information from text, allowing better decision-making. (Wassermair, 2023)

Emotion detection is one of many applications of Natural language processing (NLP). It combines computational linguistics with machine learning models to make it possible for computers to understand, generate and recognize speech and text. Moreover, emotion detection is part of the NLP branch Sentiment Analysis which tries to extract subjective qualities from the text such as emotions, sarcasm, confusion, or attitudes. (IBM, n.d.)

Paul Ekman, a pioneer in the study of emotions, identified six main emotions, happiness, sadness, fear, anger, surprise, and disgust, in a study he presented in 1971. (Ekman & Friesen, 1971) In NLP applications, categorical emotion models such as the one proposed by Ekman are commonly used to classify emotions because of their discrete categories. (Nandwani & Verma, 2021)

The media company Banijay Benelux aims to gain a better understanding of what components of their TV series are gaining attention from viewers. Together with 3Rivers, a media consultant firm, they are creating a tool for content classification in which information such as the present actors, the actions performed, or the expressed emotions are tagged by the minute and divided into fragments. To automate the emotion labeling, that is currently done by hand, they approached us to create an emotion classification model. As they already labeled the emotions of multiple seasons of the TV series Expeditie Robinson, we are using that data as the basis for our project.

The aim of our project is to identify the six key emotions as described by Ekman and Friesen (1971) for each segment of the TV series Expeditie Robinson to gain better understanding on what emotions in the segments are interesting to the viewer. The final goal of this project is to deliver a pipeline that uses speech-to-text to extract sentences from the TV series and output the classified emotion-tags for each segment. To assess the quality of our emotion classification, we are comparing the model's emotion tags with the emotion-tags labeled by 3Rivers.

2. Data selection and Exploration (EDA)

This section provides an overview of the datasets used throughout our project. It details the content of each dataset, including how they were annotated and their reliability. These datasets served as the foundation for training, testing, and exploring models for emotion analysis and recognition tasks. From large-scale Reddit comment collections to specially crafted dialogues on everyday topics, each dataset offers unique insights into emotional expression within different contexts. Additionally, the creation of generated data using OpenAI tools and the incorporation of a test dataset from our university's Data Science and AI program enriched our analysis and evaluation processes. Lastly, Expeditie Robinson data from Banijay provides a valuable real-world dataset from a reputable entertainment company, offering further opportunities for exploration and analysis.

2.1 Training Data

We used five datasets to train our models:

2.1.1 GoEmotions

This dataset comprises 58,000 English Reddit comments manually annotated with 27+ neutral emotions. Collected in 2020, it is the largest dataset used for the project. Comments were gathered from various subreddits with at least 10k comments, excluding non-English and deleted comments. Annotations were conducted using a taxonomy developed to maximize coverage and minimize overlap. Rigorous curation measures ensured dataset reliability. Annotations were validated through principal preserved

component analysis (PPCA) and clustering techniques, demonstrating dataset robustness. While subjective in nature, the dataset offers a broad spectrum of emotional expressions within conversational contexts, serving as a valuable resource for emotion analysis research. Information gained from Alon and Ko, 2021.

2.1.2 Friends Emotion

The Friends Emotion dataset comprises 12,606 lines from the TV show "Friends," annotated with seven emotions plus neutral. Collected in 2017, annotations were crowdsourced through Amazon Mechanical Turk. While annotations show low inter-annotator agreement, the dataset provides valuable insights into emotional expression in multiparty dialogues. It addresses a gap in text-based emotion detection datasets, offering fine-grained emotion annotations for consecutive utterances. Models like Sequence-based Convolutional Neural Networks (SCNN) with attention mechanisms outperform traditional approaches, achieving accuracies of up to 54% for coarse-grained emotions. However, some emotions, like peaceful and powerful, show higher confusion rates with neutral. Overall, the dataset offers a unique resource for studying emotional dynamics in conversational contexts. Information gained from Zahiri and Choi, 2017.

2.1.3 MELD Dataset

The MELD (Multimodal EmotionLines Dataset) comprises approximately 13,000 lines from 1,433 dialogues extracted from the TV series "Friends." Each line is annotated with emotion and sentiment labels and text. Collected in 2019, it addresses the lack of large-scale multimodal multi-party emotional conversational databases. Annotations were conducted by three annotators who considered both transcripts and associated video clips, achieving a Fleiss' kappa score of 0.43. The dataset provides a valuable resource for studying emotion recognition in conversations, offering insights into the importance of contextual and multimodal information. While primarily related to the domain of TV series dialogue, MELD contributes significantly to the development of conversational AI and multimodal emotion recognition systems. Information gained from Poria et al., 2018.

2.1.4 Daily Dialogue

The dataset consists of 13,118 conversations on everyday topics, manually crafted to reflect real-life communication. Created in 2017, it includes annotations for emotions and communication intentions. The sources are authentic dialogues, ensuring relevance. Annotations, added manually, aim for accuracy, though some subjectivity may be present. The dataset is reliable for studying natural language understanding and emotion recognition. It offers insights into everyday conversations and is not limited to specific topics, making it valuable for diverse research in language processing. Information gained from Li, 2021.

2.1.5 Generated Data

The dataset is created using OpenAI tools and contains short sentences labeled with the core emotions: happiness, sadness, anger, surprise, fear, and disgust. Each request produces a set containing approximately 50 sentences, designed to be easy to understand and convey different emotions clearly. Additionally, an example of a test set is provided to guide the structure of the sentences. All individual sets are merged to construct the final dataset, resulting in over 900 rows of data containing a variety of emotional expressions. Annotations for this dataset are algorithmically generated based on predefined criteria, relying on the accuracy of machine learning models. While the labeling aims for objectivity, subjective interpretations may exist. The dataset is not domain-specific, making it applicable for emotion recognition research.

2.1.6 Data Distribution

When we combined the GoEmotions, Friends Emotion, MELD Dataset, and Daily Dialogue, we sorted emotions into six main groups. Comparing the distribution of the individual emotions we can see a significant difference. Happiness stands out with almost 25000 occurrences compared to the count of disgust.
The plot below illustrates the data.

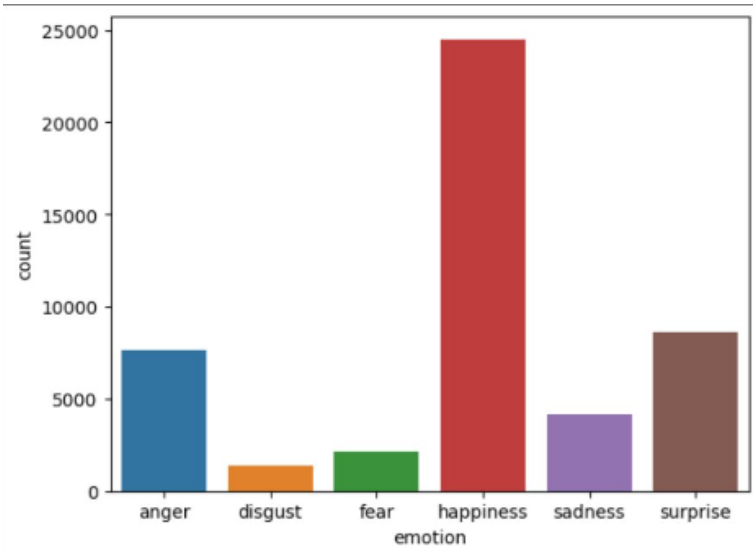


Figure 1

Data Distribution Plot of Merged Datasets

On the other hand, the Generated Data looks more even. Although happiness is still the most common, the gap between emotions is not that significant.

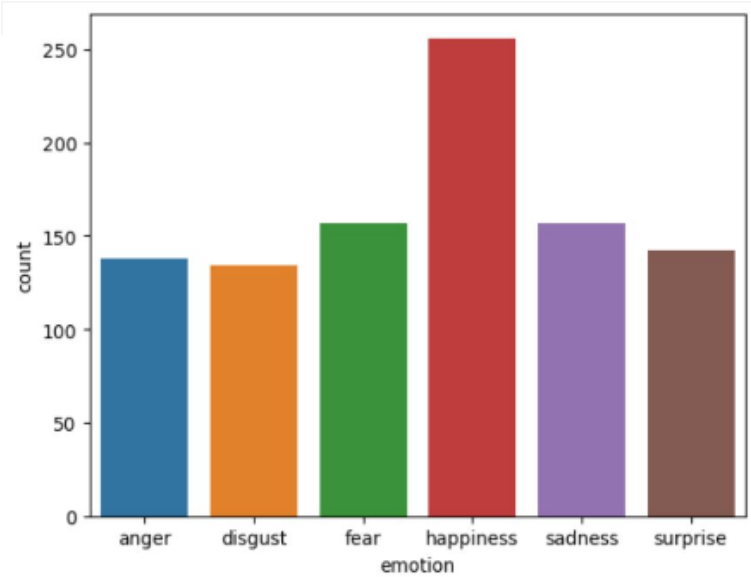


Figure 2

Data Distribution Plot of Generated Data

2.2 Test Data

The test data source is our university's Data Science and AI program, where students contributed sentences for each of the six core emotions separately. Annotations were made by the students themselves as they crafted sentences designed to express specific emotional responses. While the annotations aim for objectivity, some subjectivity may exist due to individual interpretations of emotional expression. However, the process ensured diversity and language appropriateness by using Qualtrics for submission. The trustworthiness of the test data can be seen as acceptable because it was created by students under the guidance of mentors. The final test dataset contains 1,435 rows of sentences and is not specific to any domain, allowing for the evaluation of model performance in emotion recognition tasks.

2.3 Expeditie Robinson Data

The dataset comes from Banijay, a major entertainment company, stored as a CSV file, holding details from 12 episodes of Expeditie Robinson. Each row includes info about actors, what they said, their emotions, how many people watched, and more. Since Banijay is a major player in the industry, the data can be considered well-annotated and reliable. It's likely they labelled everything accurately. The sources can be considered trustworthy due to Banijay's expertise. While the annotations aim to be objective, they might also include some personal judgments to capture emotions effectively.

3. Data preprocessing and Feature Engineering

Data preprocessing and feature engineering are crucial steps in the creation of effective predictive models in the field of natural language processing. The quality of the input data has a great impact on the performance and the output of the algorithms. Finding patterns in the data and extracting relevant information are made possible in large part by feature engineering. The data preparations and the foundation for the created predictive models are covered in this part.

In the initial phase of the predictive pipeline, the raw text data goes through a cleaning process, which includes the removal of unnecessary characters, fixing apostrophes and removing newline characters. The team performed multiple iterations of different models and feature engineering techniques. Some models were trained on a matrix of extracted features (such as Named Entities, Part of Speech tagging (POS) and sentiment) and tokenized with Term Frequency-Inverse Document Frequency (TF-IDF) in an attempt to find the best combination. TF-IDF refers to calculating the importance score of each word in the text data. This is done by multiplying two factors, the amount of times that word appears in the document and the inverse document frequency of the word across a set of documents. The named entities recognition is a technique which classifies the recognised words into predefined categories such as locations, companies, events, products etc. POS tagging involves assigning specific grammar categories (such as nouns, verbs, adjectives, adverbs, pronouns, etc.) to every word in the document. This process provides insights into the grammatical structure of the text which leads to a better understanding of word relationships and meanings.

After exploring various models and feature engineering techniques, the team ultimately decided to utilize the pre-trained BERT tokenizer for its advanced capabilities and implemented it in both of the best models in the Kaggle competition. By using its multi-layer bidirectional Transformer architecture, BERT extracts contextual information from both the left and right contexts of the input text. This enables it to generate embedding for each token, capturing complex syntactic patterns and semantic structures found in the text. As a result, the embeddings function as feature representations that contain a variety of linguistic properties including syntax, word semantics and others. Furthermore, this tokenizer uses a subword-based tokenization. That means that the unknown words are split into smaller words or characters, for the model to find meaning in the tokens. The tokenizer maps each token to its corresponding index in the vocabulary of

the chosen BERT model. Also, it marks the marking of beginnings and ends of sentences, padding and unknown tokens with special ones. Bert's ability to handle out-of-vocabulary words ensures robustness and enhances its generalization capabilities. In the case of one of the models, an uncased Bert model is used, which means that additional steps (such as lowercasing the text and applying other normalisation techniques) are applied. It's important to note that in one of the two best models, the prepossessing function uses the Stanford Sentiment Treebank corpus. That is a dataset introduced by the researchers at Stanford University and it's derived from the original SST dataset, which contains sentences from movie reviews along with sentiment labels.

4. Model Selection and Implementation

Multiple iterations were made following the recent trends and innovations in Natural Language Processing to ensure that the client's requirements were met. Various metrics including accuracy, F1 score, computational efficiency, and performance complexity were considered during the selection of the optimal model, which was determined through testing across datasets of varying sizes, origins, and characteristics as outlined in the data preprocessing section of this report.

4.1 Model Selection

Three different types of Machine Learning and Deep Learning models were used during this research:

4.1.1 No context capturing models

Models like Naive Bayes and Logistic Regression provide benefits such as simplicity, effectiveness, scalability, resilience to noise, and adaptability across various datasets (Naik, 2023). They were used to establish baseline performance on each dataset, aiming to argue whether the increased complexity of context-aware models yields significant enhancements, considering the balance between complexity and interpretability.

4.1.2 Context capturing models

Neural network architectures such as RNNs, CNNs, MLPs, LSTMs, and Bidirectional LSTMs ("A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU," 2023) were employed because they excel at capturing intricate patterns, managing sequential data, acquiring hierarchical features, and conducting automatic feature extraction, contrasting with non-context-based models considering the tradeoff between their interpretability and the interpretability of multi-attention head models

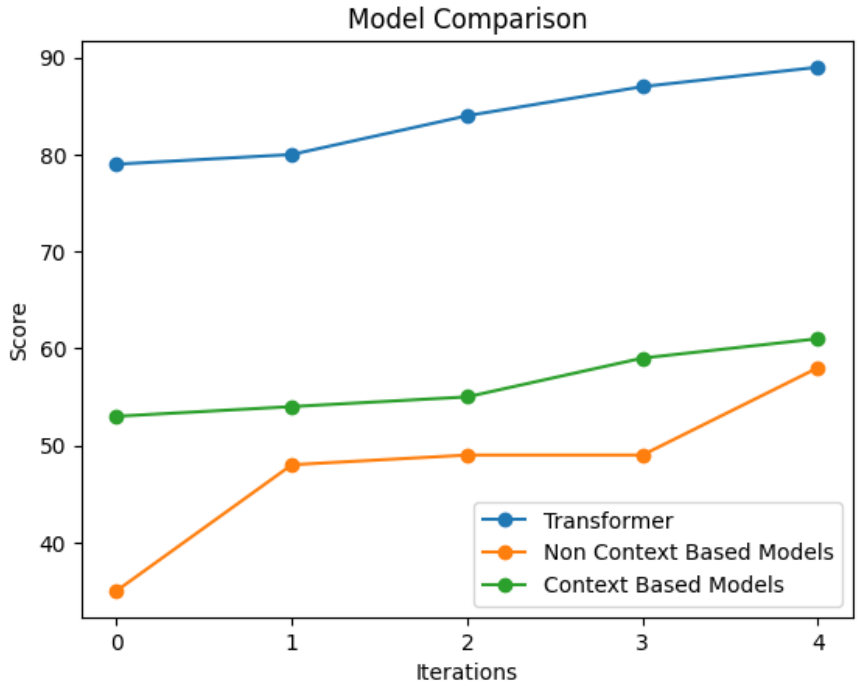
4.1.3 Multi-head attention models

Multi-head attention models, like those found in the Transformer architecture, have an advantage over the non-context and simple context-capturing models because they can capture multiple aspects of relationships within data simultaneously (Vaswani et al., 2023). This parallel processing enhances the model's ability to understand both local and global patterns.

4.4 Performance Comparison

Figure 3

Illustrates the results of the three different types of Machine Learning and Deep Learning models.



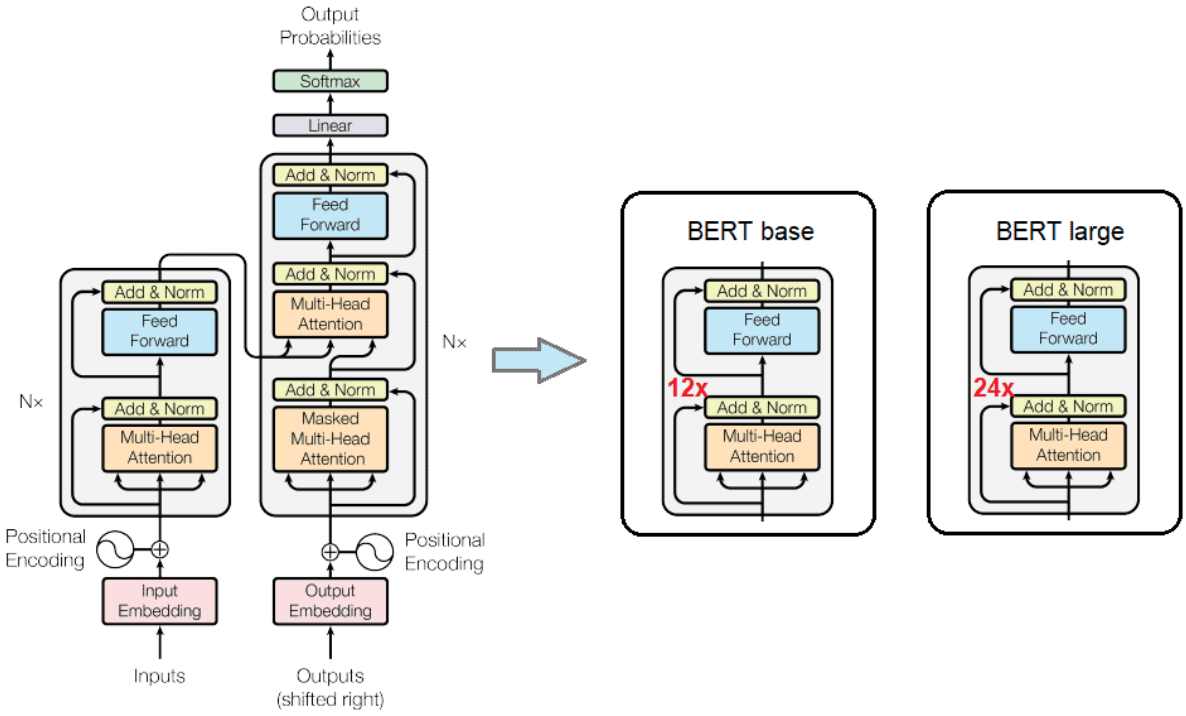
When comparing the performance of the three types of Machine and Deep Learning models, we observe that non-context-based models perform similarly to context-based models, both achieving around 50% accuracy. This accuracy is significantly higher than the random guess accuracy of 16%. However, the Transformer models demonstrate remarkable performance, averaging 84% accuracy and reaching a maximum of 90%. Based on these experimental results (figure 3), we have chosen the Transformer model as the most effective architecture for our needs.

4.5 Used Transformer models

The BERT (Bidirectional Encoder Representations from Transformers) model structure is based on the Transformer architecture and comprises multiple encoder layers (Muller, 2022). It utilizes tokenization to break input text into sub-word units and is pre-trained using unsupervised learning tasks like the Masked Language Model (MLM) and Next Sentence Prediction (NSP). BERT employs bidirectional context and fine-tuning for downstream tasks, making it a powerful tool for natural language processing.

Figure 4

The figure illustrates the difference between BERT-base and BERT-large architecture.

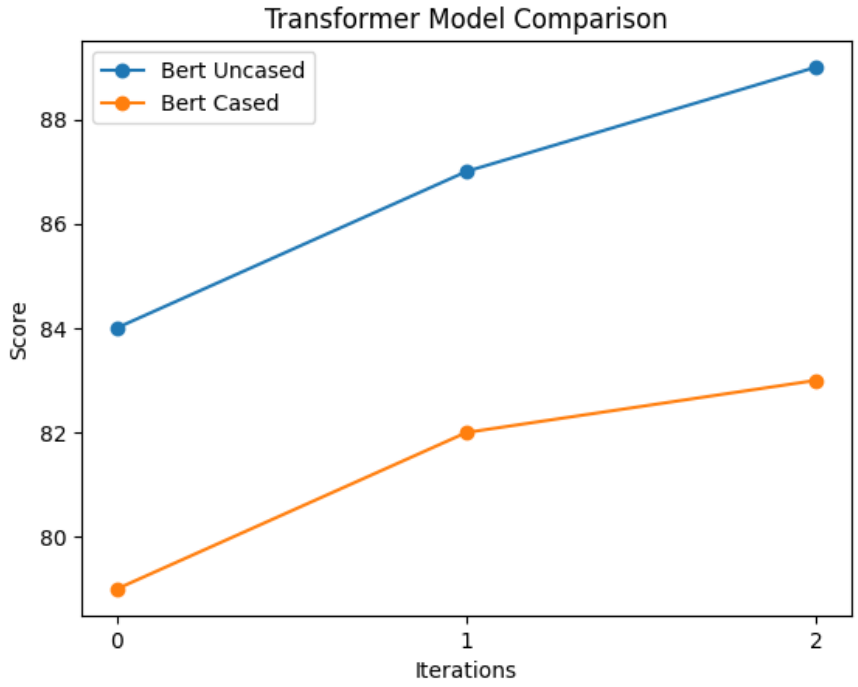


Two distinct architectures were used with two pre-trained feature extraction techniques, leading to the creation of four distinct BERT transformer models (4). BERT Base Cased and BERT Base Uncased employed a simpler architecture with fewer trainable parameters compared to BERT Large Cased and BERT Large Uncased (Table 1). Surprisingly, our results diverged from the expectation that preserving capitalization would

enhance performance by providing more contextual information for the model to learn from. Instead, the model became more confused, resulting in poorer performance, as illustrated in (figure 5).

Figure 5

Visualizes the results of Bert Base and Large grouped into two categories: Cased and Uncased.



4.6 Dataset Performance

Multiple iterations were made using the BERT base to accurately detect the emotions in the given sentences, underscoring the significance of the input data as the primary factor influencing the range of results observed in the experiments. The experiments were made using the same model architecture trained for the same number of epochs on the same data.

Transformers models are usually known for their requirement of a large sample data so that they can learn complex language patterns effectively. The large dataset helps them generalize well across different contexts, improve robustness, and avoid overfitting.

In this project, surprisingly, the smaller dataset outperformed the larger one (figure 6), contradicting typical transformer behaviour ("How do Transformers work?" 2020). This discrepancy raises concerns about the quality and composition of the test data used for comparison.

Figure 6

This chart visualizes and compares the importance of the used dataset.

Ultimately, the Bert Base architecture was chosen because the results did not demonstrate any notable differences between the Base and the Large models' architecture, likely due to limitations in the available data. Therefore, considering its faster training and prediction times, we opted to integrate the Bert Base Uncased model.

5. Evaluation Metrics and Results

After deciding on the best model, we are able to run the entire pipeline; Loading the Expeditie Robinson Episodes, transcribing and translating them into English, using our model to predict the emotions for each fragment and saving the predictions in a CSV file. To evaluate the model performance, we are conducting an extensive error analysis.

Table 1

This table compares the two different BERT Transformer architectures used in this project.

Aspect	**BERT Base**	**BERT Large**
Parameters	110M	300M
Computational Resources	Requires less resources	Requires more resources
Performance	Generally lower compared to BERT Large	Generally higher compared to BERT Base
Fine-tuning	Can provide good results	Often achieves higher performance
Use Cases	Resource-limited scenarios	Demanding applications requiring top performance

5.1 Model Performance and Error Analysis

In the first step, we compared the predicted emotions to the emotions labelled by 3Rivers to evaluate the accuracy of our model. We achieved a prediction accuracy of 53%, a recall of 53%, and an F1 score of 49%.

Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) among all instances. Recall measures the proportion of correctly identified positive instances (true positives) among all actual positive instances (true positives + false negatives). Given that the model accuracy is the same as the model recall, the false positives and false negatives might be balanced out by true positives and true negatives in such a way that the overall accuracy is the same as the recall even though they differ for each emotion. There are two possible causes for that. One cause could be that while some emotions have higher false positives or false negatives, they are balanced by other emotions with fewer false positives or false negatives, causing the accuracy and recall scores to be the same. Additionally, if one emotion has a significantly higher sample in the dataset and is well predicted by the model, it might influence the overall accuracy

and recall to be the same.

The F1 score indicates the overall effectiveness of the model in terms of both precision and recall. Based on our use case, it is therefore the most important metric to focus on, as it shows the model's effectiveness in both identifying true positives and minimizing false positives and false negatives. For our use case, all emotions are equally as important which makes it necessary to focus on having a balanced performance across different emotions. The F1 score of 49% indicates that our model struggles with precision and recall for certain emotions but overall has a moderately good performance.

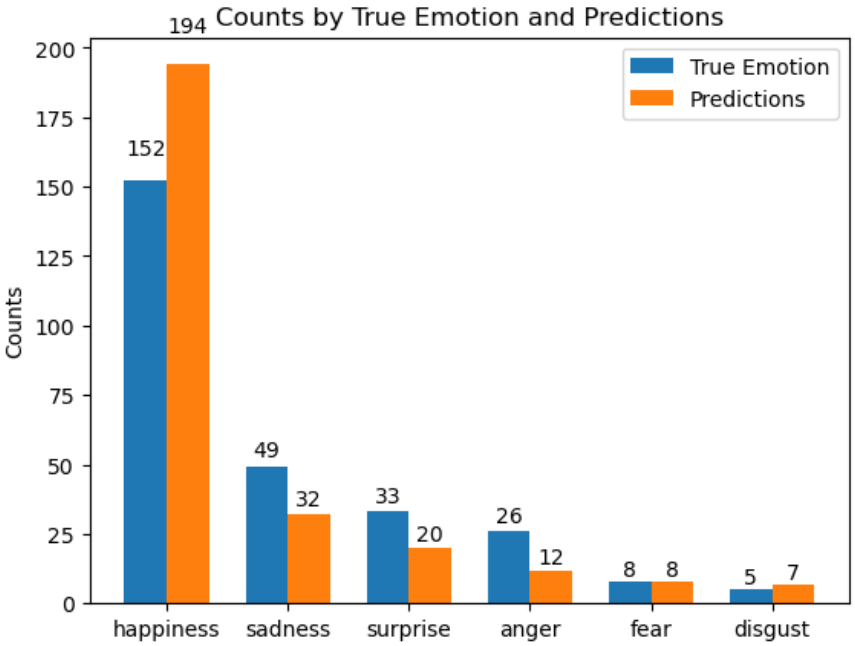


Figure 7

Number of True Emotions and Predictions.

In figure 7 we can see the distribution of the labelled emotion (True Emotion) classes and the predicted emotions (Predictions) in the Expeditie Robinson data. The graph clearly shows that our model tends to predict happiness more than other emotions. Overall, the emotion happiness is significantly more present than all other emotions combined causing the error analysis to be less indicative for all emotions with only a few sentences.

To further visualise the prediction performance of our model, we can use a confusion matrix.

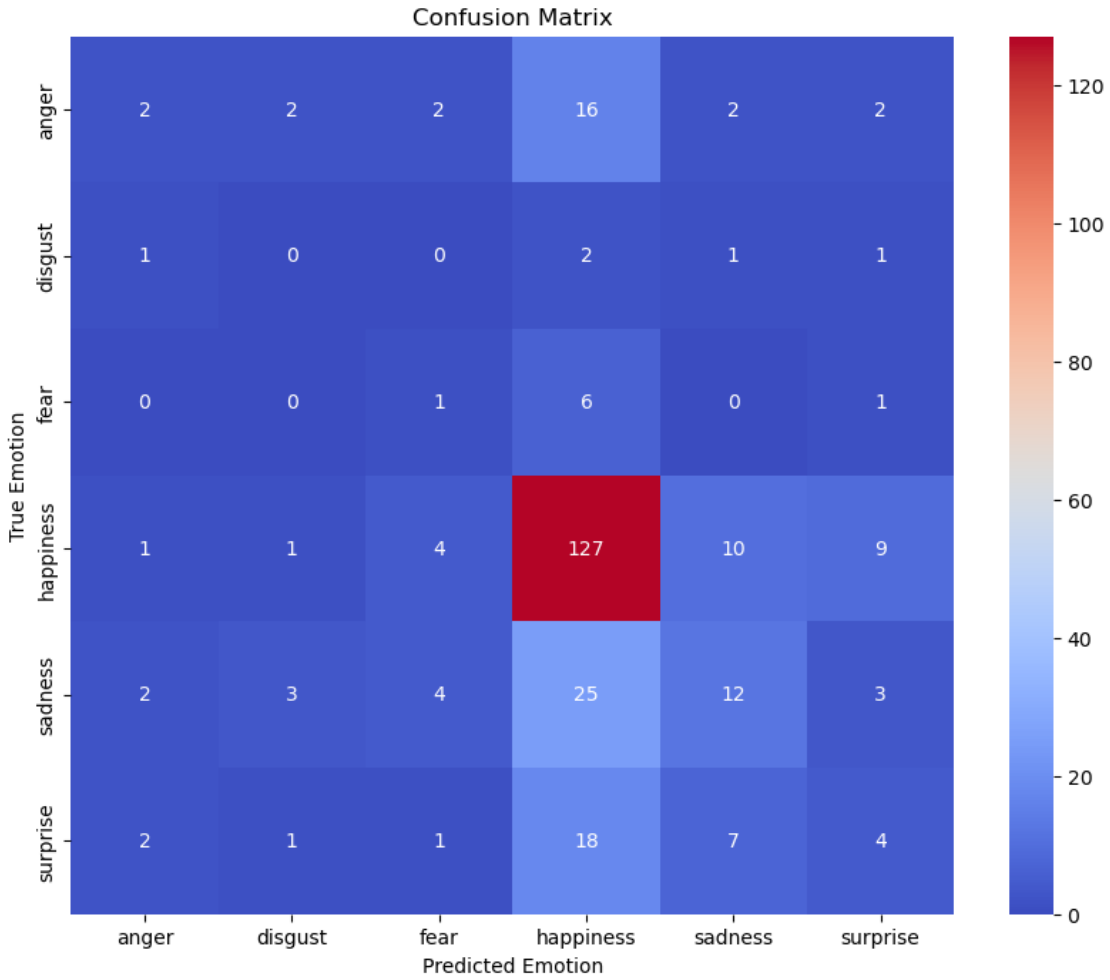


Figure 8

Confusion Matrix to visualize prediction performance.

In figure 8 it is clearly visible that happiness has by far the most correct predictions with 127. The worst performing class is disgust, with no correct predictions. Anger has 2 correct predictions, fear has 1, sadness has 12 and surprise has 4. The total false positives and false negatives for each emotion are visualized in table 2.

Emotion	False Positives	False Negatives
Anger	6	24
Disgust	7	5
Fear	11	7
Happiness	67	25
Sadness	20	37
Surprise	16	29

Table 2

Emotion Detection False Positives and Negatives

5.1.1 Word Type Analysis

To evaluate if there is a word type that occurs often that throws the model off, we conducted a word-type analysis. We evaluated the frequency for each word type for the false predictions as well as for the correct predictions. The results of the word-type analysis showed, that there was no significant difference in the word-type distribution for each emotion between the false predictions and the correct predictions.

5.1.2 Fragment Length Analysis

In the next iteration of our error analysis, we conducted a fragment length analysis to evaluate if the length of a fragment influences the accuracy of the predictions. The fragment length analysis is visualized in figure 9.

Since only happiness has a lot of fragments, the distribution of the bars for the other emotions is not as meaningful because there are not enough true and false predictions to compare to each other. When looking at the true and false predictions for happiness, the distribution of the fragment length is very similar. Therefore, the fragment length has an insignificant influence on the prediction accuracy.

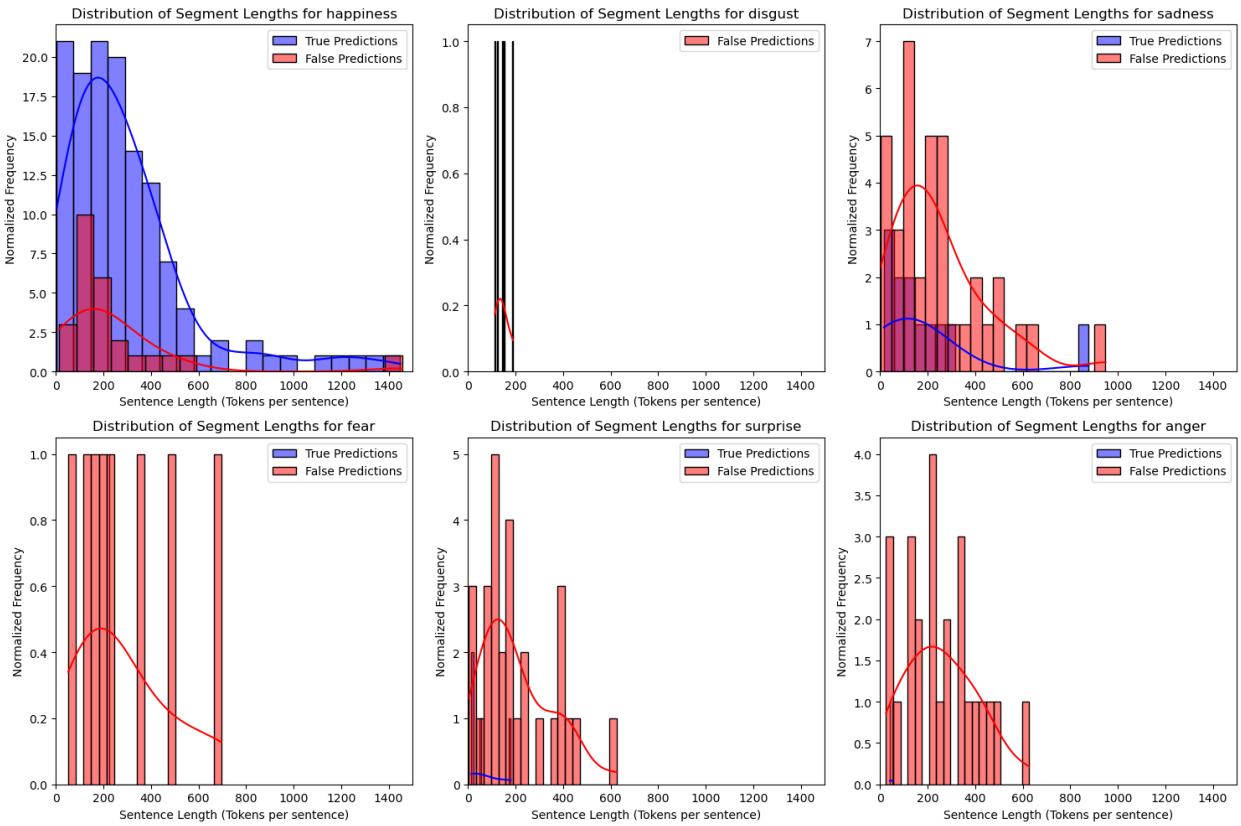


Figure 9

Fragment Length analysis for each emotion

5.1.3 Text Feature Importance Analysis with LIME

In the final step of our error analysis, we conducted a text feature importance Analysis using the Explainable AI tool LIME. We applied LIME to fragments that were predicted wrong by the model. We were aiming to find out which words are causing the model to predict the wrong class and if we can find any pattern. Figure 10 is an example of our LIME Analysis on a fragment with the true labels sadness and anger that was predicted wrongly by our model.

In figure 10 it is visible that the model did not use many words to predict the emotion of the text. This is due to the model being trained on sentences with connected words and phrases that it could not find in the segment. The fragment that the model is trying to classify does not have a sentence structure which can lead to missing context and

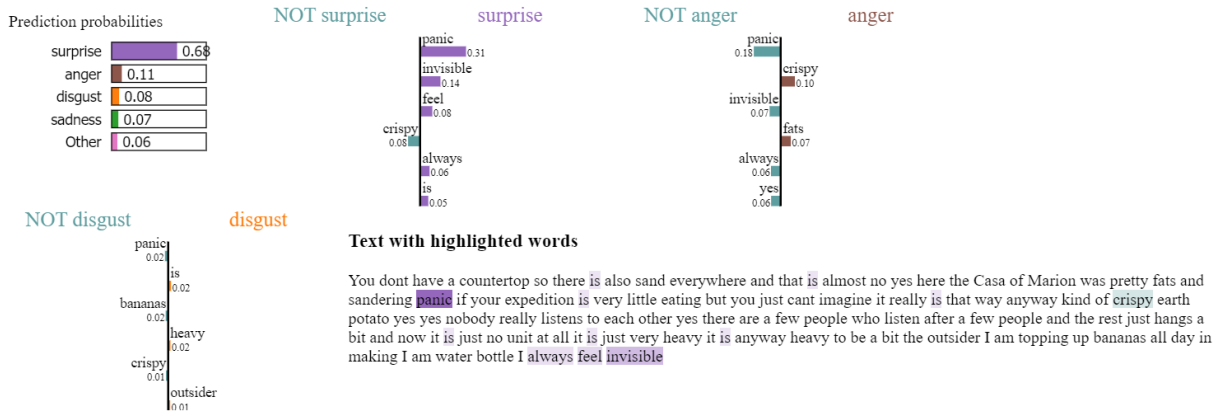


Figure 10

LIME Analysis of a fragment with the true labels: sadness and anger

meaning for the model. As most of our translated sentences have no punctuation or sentence structure, it explains why our model predicts so many emotions wrongly.

5.2 Strengths, and Limitations

Based on the model's performance and the results of the error analysis, we identified several strengths and limitations of the model when it comes to predicting the emotion for each fragment of the show Expeditie Robinson.

5.2.1 Strengths

One of the main strengths of our model is the utilisation of BERT as our base model, as it is easy to scale and has various options to fine-tune the model to the exact requirements of the use case.

Another strength of our model is, that it is performing with an F1 score of 89% on the test data. This means, that our model is performing very well on data with the same structure as the data that was used for training.

5.2.2 Limitations

Even though our model is performing well on the test data, there are some limitations when it comes to applying the model to our use case. One of the main issues is, that the model itself was trained on sentences and not on fragments of text which means

that it performs very well on shorter sentences but lacks accuracy when predicting emotions for a whole paragraph of text.

Another limitation of the model is the lack of domain-specific data as input for the model training. This causes the model to lack training on words that are more specific to the show Expeditie Robinson.

Furthermore, the model was trained on English text, but the show is Dutch and needed to be translated before we could apply the emotion detection. The translation caused a lot of context and implicit emotions to get lost, which ultimately influenced the emotion prediction as the sentences in each fragment don't have an English sentence structure and lack meaning in many cases.

A final limitation of the model is the subjective ground truth that was provided to us by 3Rivers. They labelled the emotions for each fragment based on subjective themes and emotions that might not always be represented in the transcription of the show.

6. Future Recommendations

Based on the limitations that we identified, we propose the following future steps:

The most important step to increase the model's performance is to create a better dataset for the model to be trained on. This includes gathering more data and ensuring that the labelled sentences are longer, or ideally, labelled in paragraphs. It is also important that terminology and common phrases of the show are included in the dataset.

We also recommend considering training the model solely on Dutch data to avoid issues with translation. This will enable the model to learn Dutch language structure and associate common Dutch phrases for each emotion. To further increase the model performance, we recommend using the Dutch BERT model RobBERT as it provides the same benefits of using BERT as a base model while being trained specifically for Dutch NLP tasks.

Finally, we encourage labelling more episodes of the show while ensuring that the labelled emotions apply to the spoken text and not the emotions of the viewer while

watching. If more data from the show is labelled, it can be used for model training to increase the model performance.

7. Discussion and Conclusion

We developed an emotion classification pipeline for Expeditie Robinson using advanced NLP techniques like BERT. Our model achieved 53% accuracy, excelling in identifying happiness but struggling with emotions like disgust and surprise. Error analysis highlighted areas for improvement, including word type influence and text feature importance. Challenges included the dependence on translated text and limited domain-specific data. To enhance performance, we recommend gathering more specific data, training on Dutch language data using models like RobBERT, and refining the labelling process for accuracy. This project showcases NLP's potential in automating emotion classification for media content, with future work aimed at advancing emotion analysis pipelines for entertainment applications.

References

Alon, D., & Ko, J. (2021, October). *Goemotions: A dataset for fine-grained emotion classification*.

A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. (2023).

Ekman, P., & Friesen, W. V. (1971). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

How do transformers work? (2020).

IBM. (n.d.). What is natural language processing? [https://www.ibm.com/topics/natural-language-processing [Accessed: 02.04.2024]].

Li, Y. (2021). Dailydialog [Available at http://yanran.li/dailydialog].

Muller, B. (2022, March). *Bert 101 state of the art nlp model explained*.

Naik, S. (2023, March). Naive Bayes vs Logistic Regression. https://www.educba.com/naive-bayes-vs-logistic-regression/

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Wassermair, P. (2023, April). *Emotion classification of text* [Available at https://epub.jku.at/obvulihs/download/pdf/8620963?originalFilename=true].

Zahiri, S. M., & Choi, J. D. (2017). Emotion detection on tv show transcripts with sequence-based convolutional neural networks.