UNDERSTANDING EMOTIONS IN TEXT

Mauro van Hulst, Imani Senior, Hubert Waleńczak, Wojciech Stachowiak

1 INTRODUCTION

Understanding and interpreting human emotions expressed in textual data is crucial in the modern digital age for a variety of purposes, from social media engagement and mental health monitoring to marketing strategies and customer sentiment analysis. As an aspect of natural language processing (NLP), emotion recognition aims to interpret the complex nuances of human emotional expression captured in written communication. This paper provides a thorough examination of the approaches, difficulties, and learnings from creating reliable emotion detection models.

The process of recognizing emotions is complex and includes several steps, including gathering data, preprocessing, feature engineering, choosing a model, and testing it. Each stage presents unique challenges, requiring careful consideration and methodological rigor to achieve accurate and reliable results. Effective emotion recognition systems are built on an understanding of the complexities of human emotional expression, linguistic variability, and contextual subtleties.

First, we conduct a comprehensive analysis of emotion-labeled datasets obtained from a variety of sources, such as scripted dialogues, literary works, and social media platforms. These datasets offer an extensive collection of emotional expressions recorded in a variety of linguistic contexts, which is extremely helpful in understanding the intricacies of human emotions. Robust emotion recognition models are built upon the patterns and trends that are discovered through painstaking data processing and investigation.

The choice and improvement of feature engineering tactics and text pre-processing methods specifically designed to extract significant insights from textual data are essential to our investigation. Each stage that is taken affects the effectiveness and generalize ability of the resulting emotion recognition models, from tokenization and normalization to the extraction of linguistic features and sentiment indicators.

The difficulties and factors we come across during the process of selecting a model, feature engineering, and data preprocessing all have an impact on the creation of efficient emotion recognition systems. Every choice we make, from choosing the best machine learning algorithms to assessing the performance of the model, helps to improve and optimize our methodology.

2 Data Processing and Exploration

2.1 Training Data

Our data processing and exploration task uses a diverse combination of different emotion-labeled datasets as training data. These datasets each provide unique insights into how human emotions are expressed in various contexts. The wide range of sources and annotation methods represented in these datasets adds to the combined dataset's complexity and depth.

- **GoEmotions**: This dataset offers an extensive collection of real-world textual data reflecting a wide range of emotions expressed within online communities. It consists of 58,000 manually annotated English Reddit comments. There is a broad range of language styles, topics, and emotional expressions in Reddit comments because of their diverse nature.
- **SMILE Twitter Emotion Dataset**: This dataset, which contains 3,085 tweets mentioning the Twitter handles of British museums, provides an insight into the emotional reactions that these cultural institutions generate on social media. With limited character constraints, emotion recognition models encounter a unique challenge of recording subtle variations in emotional expression due to the shortness of tweets. It's worth noting that there has been feedback suggesting that the performance of models trained on this dataset is not as effective as expected. Some findings indicate that the dataset might not significantly contribute to improving performance, with models yielding similar results even with random inputs. Further investigation may be necessary to understand potential issues or irregularities within the dataset that could affect model performance.
- **Friends Emotion-labeled Dialogues and MELD**: These datasets provide scripted conversational data that reflects a blend of drama, humor, and personal relationships. They are composed of manually annotated utterances from episodes of the popular TV show Friends. The MELD dataset has been expanded and improved, providing a larger and more varying database of conversations for the purpose of testing and training emotion recognition algorithms.
- **CARER**: The CARER dataset is an extensive collection of social media data that represents a wide range of emotions expressed in online discourse. It consists of over 400,000 English tweets that were gathered through noisy labels and annotated through remote supervision. The enormous amount of data provides valuable insights into the distribution and patterns of emotional expression on social media platforms, even though the distant supervision approach may introduce noise.
- **Affective Text (Test Corpus of SemEval 2007)**: This dataset, comprising 250 newspaper headlines collected for the SemEval 2007 task, offers an organized set of textual data with gold labels assigned for emotion classification. The dataset's controlled structure makes it easier to benchmark and assess emotion recognition models in a consistent environment.
- **Daily Dialogue**: The Daily Dialogue dataset offers an abundance of conversational data that reflects everyday interactions and emotional dynamics, with over 13,000 manually labeled dialogues about topics related to daily life, annotated with topic, emotion, and communication intention.
- **EmoBank**: Comprising 10,000 sentences annotated according to expressed and perceived emotions, with mappings to Ekman's 6 Basic Emotions, the EmoBank dataset offers a nuanced perspective on emotional expression in written language, capturing both the author's intended emotion and the reader's perceived emotion.
- **Affect Data**: This dataset, which was assembled and annotated from 185 fairytales, has more than 15,000 sentences categorized by mood and emotions. Fairytales offer a structured narrative framework for examining how emotions are portrayed in literature and storytelling, despite their seeming disconnection from real-world interactions.

2.2 Exploration

We performed exploratory data analysis which allowed us to further understand the data we were working with. Visualizing the frequency of labels and generating word clouds from the datasets enabled us to select the most fitting datasets, and uncover underlying issues that could affect the performance of our models.

We found that distribution of feeling labels in all datasets that we gathered is very uneven with emotion of happiness dominating in terms of instances count (**see Figure 1**). On the other hand the word cloud of the dataset shows that, with exception of names, the most common words can be connected to emotions of a person which is important for models that use bag-of-words approach (**see Figure 2**).

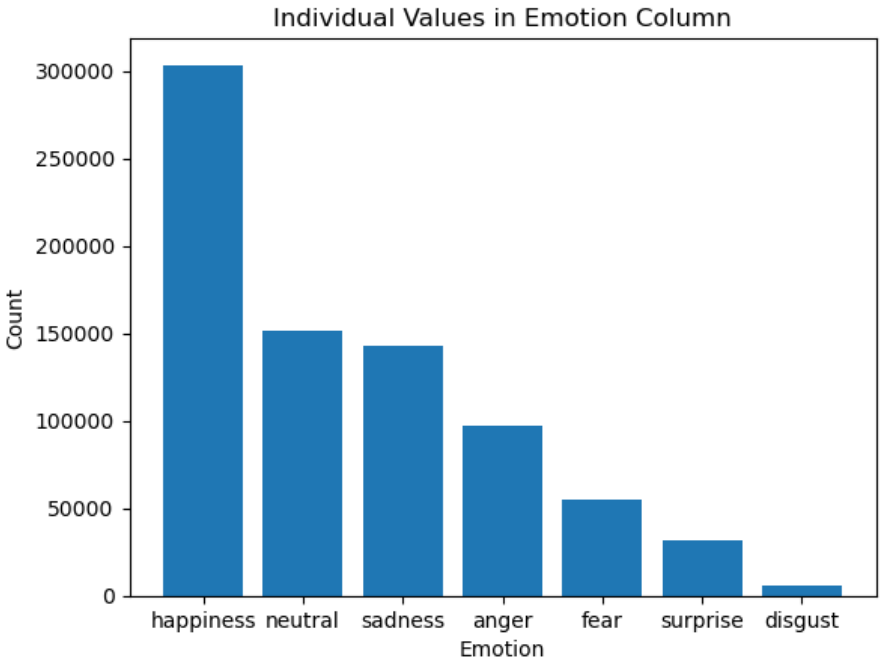

Figure 1. Distribution of emotions in all datasets collected.

To insure high quality of data we utilized Naive Bayes algorithm to test all combinations of datasets and separated the two best dataset, as they returned best results, to utilize and looked closer at them, CARER and Go Emotions, as well as some of the worst datasets such as SMILE.

The SMILE dataset is quite small compared to the combined dataset. However it is extremely skewed towards happiness with combined count of other emotions being slightly over 10% of the dataset(**see Figure 3**) with most popular words being mostly disconnected from emotion of a person that are not often used in television shows and as such should be avoided(**see Figure 4**).

CARER dataset on the other hand turns out to have fairly balanced distribution of labels compared to other datasets with happiness having only 50000 more examples than the second most common emotion (**see Figure 5**) however it has only five out of 6 basic emotions that we are trying to predict. Additionally it is dominated by words capable of conveying emotions which is helpful for our task. All in all, this is a very promising dataset that can be combined with a different dataset to fill in the gap of missing emotion of fear (**see Figure 6**).

Another dataset that stands out, Go Emotions, is less balanced then CARER and will need some form of balancing to achieve the best results, additionally it contains seven emotions, seventh being the neutral emotion (**see Figure 7**). Names are very common in the sentences this dataset contains, however the other common words often carry emotional meaning (**see Figure 8**).

2.3 Test Data

The test data comprises sentences provided by students, each delivering five sentences for each of the six emotions: happiness, sadness, anger, disgust, surprise, and fear. This student-generated dataset offers a controlled environment for evaluating the performance of emotion recognition algorithms on a standardized set of stimuli. By soliciting contributions from multiple students, we ensure diversity in linguistic styles, personal experiences, and emotional expression, thereby enriching the test dataset and facilitating a more robust evaluation of model generalization and performance across different emotional states.

2.4 Use Case Data (Expeditie Robinson)

The use case data involves speech-to-text transcriptions from episodes of Expeditie Robinson, a popular reality television show. For this use case, predictions will be made using the dataset created from these episodes, allowing us to analyze the emotional dynamics and expressions within the context of real-world interactions and scenarios depicted in the show.

Figure 2. Word Cloud of all words in all datasets collected.

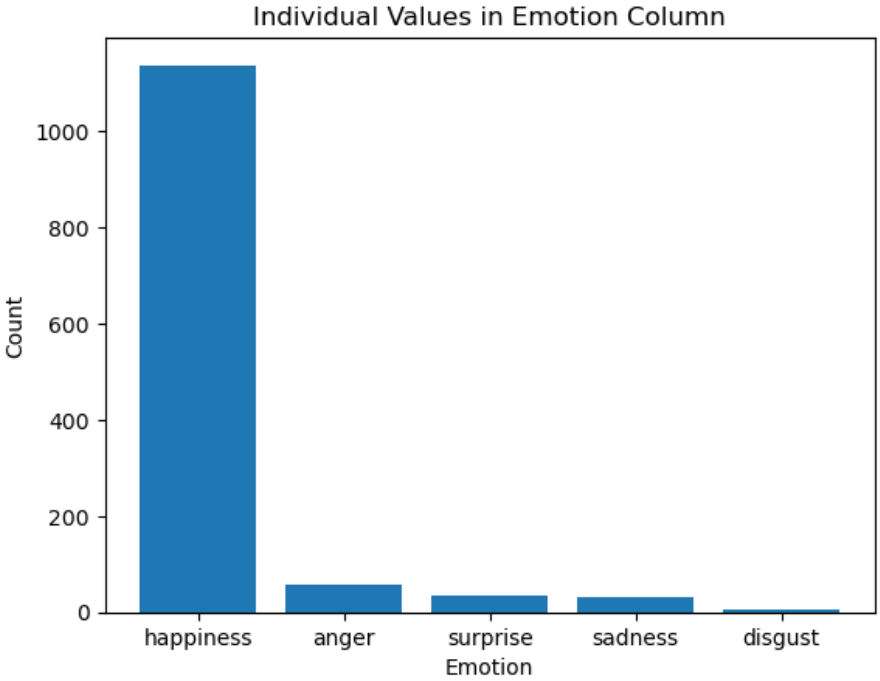

Figure 3. Distribution of emotions in SMILE datasets collected.

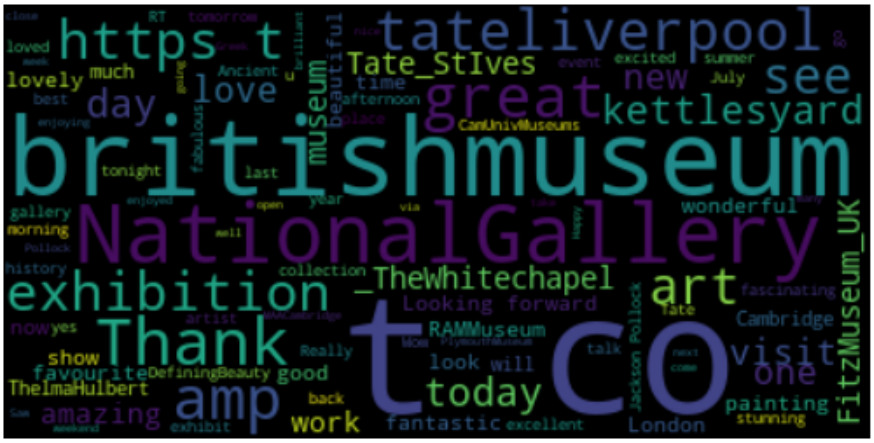

Figure 4. Word Cloud of all words in SMILE datasets collected.

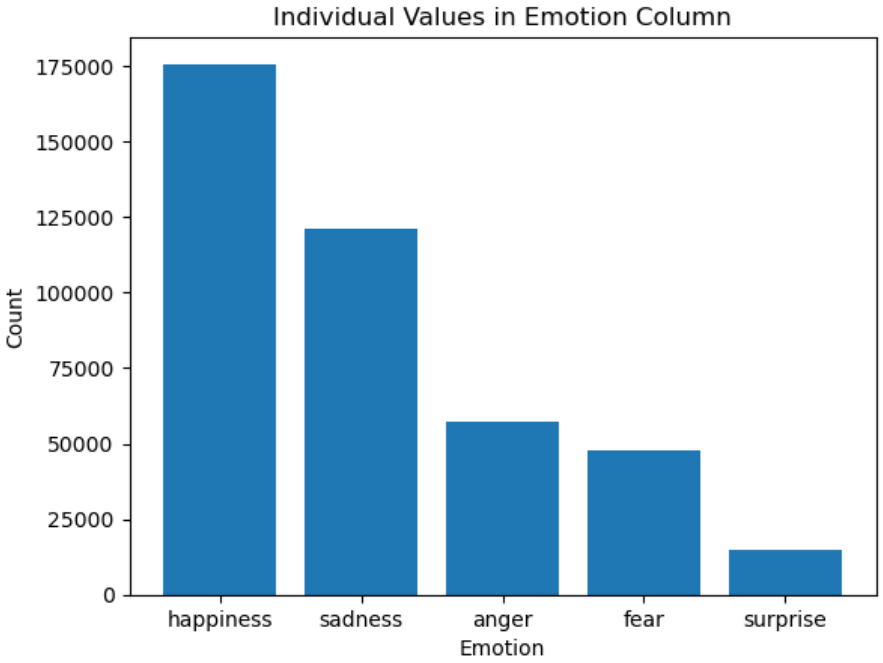

Figure 5. Distribution of emotions in CARER datasets collected.

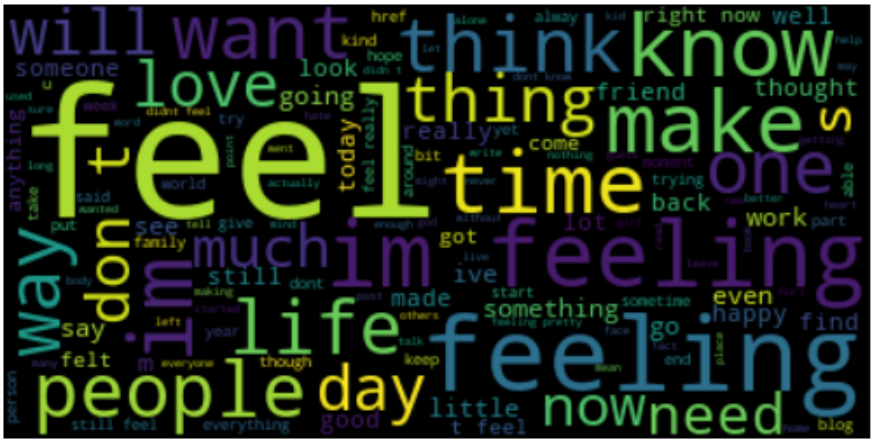

Figure 6. Word Cloud of all words in CARER datasets collected.

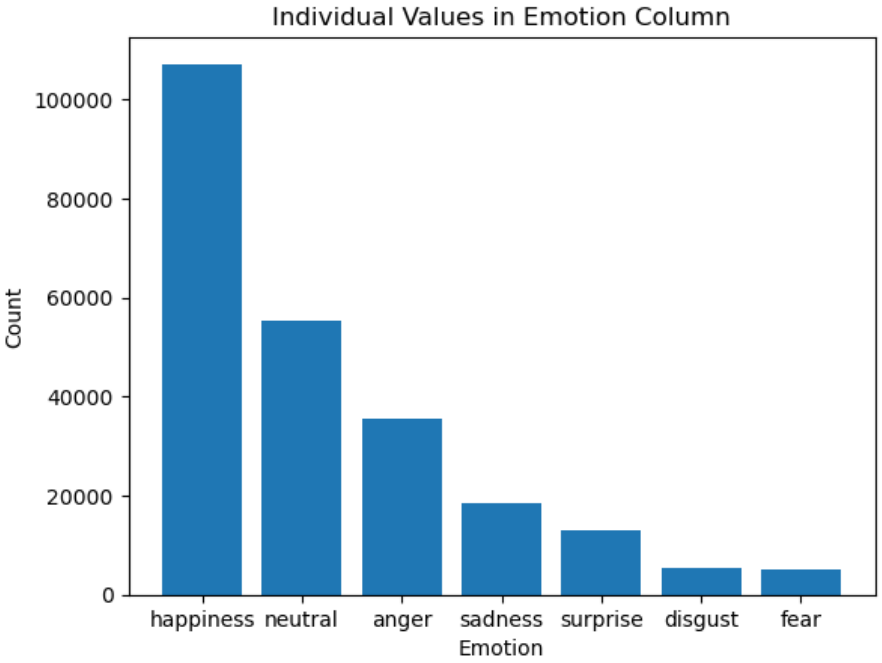

Figure 7. Distribution of emotions in Go Emotions datasets collected.

Figure 8. Word Cloud of all words in Go Emotions datasets collected.

By leveraging speech-to-text technology, we can capture the nuances of spoken language and vocal cues, enhancing the fidelity and realism of the dataset for emotion recognition tasks.

2.5 Data Annotation and Reliability

The datasets used in our data processing and exploration task vary in annotation methods and reliability, each offering unique strengths and considerations for downstream analysis and model training:

- **Manual Annotation**: Datasets such as GoEmotions, SMILE, Friends, MELD, and Daily Dialogue are manually annotated, ensuring a high level of reliability and accuracy in emotion labeling. Human annotators meticulously label each instance, providing rich contextual information and nuanced interpretations of emotional expression.
- **Distant Supervision**: The CARER dataset is annotated via distant supervision, where labels are inferred from external sources or heuristics, potentially introducing noise or inaccuracies. While distant supervision enables annotation at scale, it may require additional pre-processing and quality control measures to ensure the reliability of annotations.
- **Gold Labels**: The Affective Text dataset has gold labels assigned, indicating a high level of reliability and confidence in the annotated emotions. Gold-standard annotations serve as a benchmark for evaluating the performance of emotion recognition models and comparing results across different datasets and approaches.
- **Perceived Emotions**: EmoBank incorporates annotations for both expressed and perceived emotions, offering a multi-faceted perspective on emotional expression in written language. By capturing the emotions perceived by readers in addition to the emotions expressed by authors, EmoBank provides valuable insights into the interpretation and reception of emotional content.
- **Unique Perspectives**: Affect Data's annotation from fairy tales provides a unique perspective on emotional expression in storytelling and literature. While not directly reflective of real-world interactions, fairy tales offer a structured narrative framework for exploring the portrayal of emotions and mood in fictional contexts.

Overall, understanding the annotation methods and reliability of the datasets is essential for interpreting the results of data processing and exploration tasks and ensuring the validity and generalizability of findings.

3 Data Pre-processing and Features

To get the most out of gathered datasets, we performed pre-processing and feature extraction. Those are the steps that sped up the model training process later on, as all the possible features were ready to use if we decided they might be useful to the currently trained model.

The features were extracted using the Python NLP library - SpaCy, and its extensions: SpaCyTextBlob (adaptation of TextBlob for SpaCy pipe method), and spacy-ngram (SpaCy tool to extract n-grams from text). The table with all

extracted features and level of extraction (per token or per document) can be seen below. When chosen, those features were later used as input for the models. Those features are used to compliment the main input which is the sentence itself.

Since the variety of models created later require different kinds of pre-processing, there was no major pre-processing done. We decided to work on multiple datasets managed individually, so the state of processed data differs. From leaving the sentences as a raw version of what was acquired from the datasets, to discarding punctuation, stop words, and leaving only ASCII characters to ensure the usefulness, given that the datasets are rather noisy. It is difficult to prepare the bare sentence for modelling process because of the differences in approach for each model type. We used count vectorizers to create bag of words representations later used in both statistical models and neural network approaches such as RNNs. Some statistical models use tf-idf encoding to get a more detailed information about the sentences. We also implemented vector embedding using both vectors pre-trained on GoogleNews dataset and vectors we trained using our dataset. Finally, BERT transormer models use its own vectorizer that creates additional special tokens that BERT models understand.

Feature	Level (doc or token)	Explanation
entities	document	entities tagged in this document
noun chunks	document	nouns detected to belong to one object
2-grams	document	2-grams in a given document
3-grams	document	3-grams in a given document
sentiment	both	sentiment from positive to negative (-1 to 1)
subjectivity	both	how personal given token/document is (0 to 1)
part of speech	token	part of speech recognised by POS classifier
tag	token	more detailed part of speech
lemmatized	token	lemmatized tokens
normalized	token	normalized tokens
dependency	token	token dependency in the document
tf-idf	token	tf-idf probability of a given token

4 Model Selection

The primary factor in predicting emotions in text is the model utilized. There are many different models that can be used for emotion classification. In this section I will discuss the different models that we have used, simply explain the underlying mechanics of how they work and how they compare to each other.

4.1 Naive Bayes

Thanks to its simple statistical approach, training time and resources are the smallest compared to other methods we used. By looking at probabilities of words appearing in a class and multiplying them together we achieve the probability of given text in that class and choose the class with highest probability Webb et al. (2010). The model robust statistical algorithm requires just the text to classify without any additional pre-processing. To further improve the model performance we used the term frequency–inverse document frequency (TF-IDF) to help with fact that some words appear more often in general then other. Downside of Naive Bayes is that it does not understand sentence structure as it uses bag-of-words, that is unordered collection of words.

4.2 Recurrent Neural Network and Convolutional Neural Network

There are multiple types of Recurrent Neural Network (RNN) with two of them prevailing: Long-Short-Term-Memory (LSTM) Hochreiter and Schmidhuber (1997) and Gated Recurrent Unit (GRU) Cho et al. (2014), both of them having similar results in many tasks Yin et al. (2017) and as such we decided to use LSTM layers for our RNN models. For the architecture we decided to use either one or two LSTM layers with dimensionality of the output at 128. We experimented with a couple of different head configurations of the model (Dense Neural Networks after LSTM layers) to find the best combination.

We also used Convolutional Neural Network model (CNN) as it may perform better in some cases than RNN Yin et al. (2017). For the CNN architecture, we adopted a multi-input approach, where each input type (e.g., sentence, part-of-speech tags, polarity) is processed independently through one-dimensional convolutional layers. Specifically, each input type is subjected to two consecutive convolutional layers with a kernel size of 5. The first convolutional layer outputs 128 feature maps, while the subsequent layer generates 256 feature maps. Following the convolutional layers, a one-dimensional max-pooling operation is applied to downsample the feature maps.

After processing all input types through their respective convolutional layers, the resulting feature maps are concatenated into a single representation. This concatenated feature vector is then passed through two fully connected (dense) layers, each consisting of 256 units. Finally, a single dense layer with six outputs, one for each emotion category, produces the final predictions.

Both of these architectures are more advanced than Naive Bayes and are capable of outperforming it due to their ability to capture sentence structure. However, they require more resources for training and utilization.

4.3 Transformer

Transformer models are current state-of-the-art models in Natural Language Processing thanks to utilization of self-attention layers. These mechanisms enable the model to identify the context of each word in a sentence by assigning varying degrees of importance to different words based on their relevance to each other within the context of the entire sentence. This approach allows Transformers to capture long-range dependencies and contextual nuances in text more effectively than previous architectures. Transformer models employ multiple attention heads in parallel. Each attention head focuses on different aspects of the input, allowing the model to capture various types of information simultaneously. Furthermore, Transformer models stack multiple layers of self-attention and feed-forward neural networks, enabling them to learn complex hierarchical representations of text Vaswani et al. (2017). We utilized a pre-trained BERT model, widely used transformer for NLP tasks developed by Google, with input size of 512 tokens. Additionally, we designed a custom head for predicting 6 basic emotions using the transformers library. In the finished pipeline we use another SOFTMAX layer on the output of the model to limit the output values between 0 and 1. This helps with identifying neutral emotion as it is not directly predicted in our model and we use threshold for that.

5 Evaluation, Error Analysis

The evaluation of emotion recognition models is crucial to understand their performance and identify areas for improvement. In this section, we analyze the effectiveness of different models—Naive Bayes, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and BERT—based on various metrics such as accuracy, precision, recall, and F1 score.

5.1 Model Performance Comparison

Let's start by comparing the performance of each model using the provided metrics:

Model	Accuracy	F1 Score	Precision	Recall
Naive Bayes	0.17	0.07	0.15	0.17
RNN	0.19	0.16	0.18	0.19
CNN	0.16	0.11	0.23	0.16
BERT	0.53	0.50	0.62	0.53

Table 1. Performance Metrics of Different Models

From the table, it's evident that BERT outperforms the other models significantly in terms of accuracy, F1 score, precision, and recall. Naive Bayes has the lowest performance across all metrics, indicating its limitations in capturing the complexities of emotional expression in text. RNN and CNN perform moderately, but they fall short compared to BERT.

5.2 Error Analysis

Next, let's delve into the errors made by each model to gain insights into their shortcomings and potential areas for improvement.

5.2.1 Naive Bayes

The Naive Bayes model shows low accuracy and F1 score, indicating its inability to capture the nuances of emotional expression. It tends to predict disgust predominantly, possibly due to the imbalance in the dataset or the limitations of the bag-of-words approach. The model struggles with distinguishing between emotions, resulting in poor precision and recall scores across all categories.

5.2.2 RNN

The RNN model exhibits slightly better performance compared to Naive Bayes but still falls short in accurately predicting emotions. It tends to misclassify sadness as happiness, which could be attributed to the similarity in linguistic patterns between the two emotions. While RNN captures some aspects of emotional expression, its performance is hindered by the complexity of sentence structures and contextual dependencies.

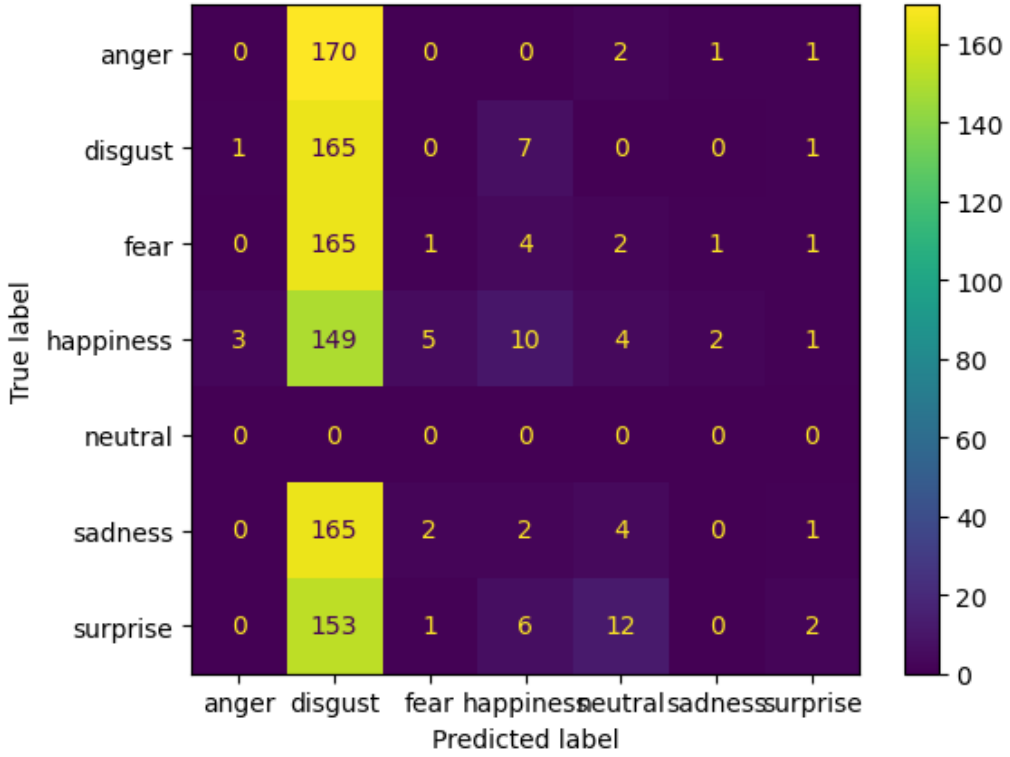

Figure 9. Confusion Matrix for Naive Bayes

5.2.3 CNN

Similar to RNN, the CNN model struggles with accurately identifying emotions, particularly sadness and anger. It shows high prediction scores for happiness but performs poorly in predicting other emotions, indicating a lack of generalization. The model's architecture may not effectively capture the contextual nuances necessary for accurate emotion recognition.

5.2.4 BERT

BERT demonstrates superior performance across all metrics, reflecting its ability to leverage contextual information and linguistic nuances. While BERT excels in predicting happiness, it also achieves relatively high scores for other emotions, showcasing its robustness and generalization capabilities. The model's success can be attributed to its pre-training on large corpora and its ability to understand complex language structures.

5.3 Comparison and Selection

Each model offers unique strengths and limitations, catering to different requirements and preferences. Naive Bayes provides simplicity and efficiency but may struggle with capturing complex sentence structures. RNNs with LSTM units excel at capturing temporal dependencies, while CNNs offer an alternative approach, capturing local patterns effectively. Transformers represent the cutting-edge in NLP and demonstrate impressive performance in capturing emotional nuances.

Ultimately, the choice of model depends on the specific requirements of the application, including available resources, desired performance metrics, and the nature of the input data. In our project, we selected models based on a balance between performance and resource requirements, ensuring optimal performance within our constraints.

6 Discussion

6.1 Client Requirements and Project Objective

Our project aims to address the needs of Banijay Benelux, a leading TV show producer, by developing a pipeline for emotion classification in the popular TV series Expeditie Robinson. With the objective of understanding viewer engagement and interest, Banijay Benelux seeks to analyze emotions expressed throughout the episodes.

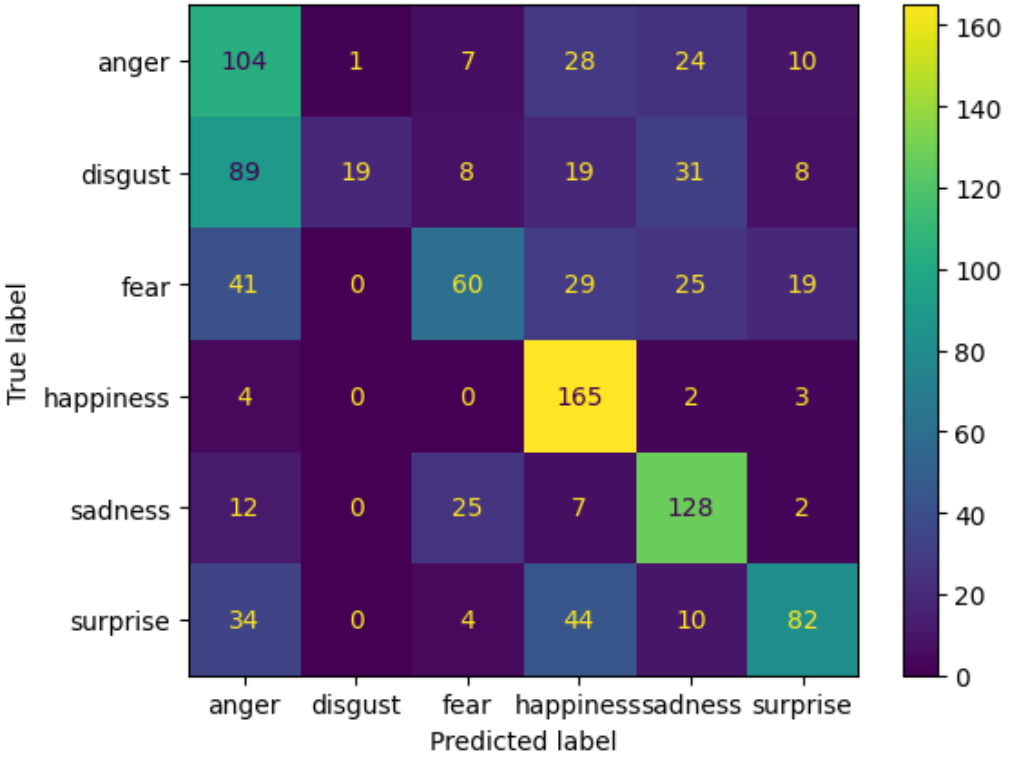

Figure 10. Confusion Matrix for BERT

Expeditie Robinson follows a structured format, allowing for the identification of emotional dynamics across various segments, from the initial arrival of candidates to the intense tribal councils and finale games. To achieve this, we have developed a pipeline that, given video episode and a table with segments separated using seconds from the episode start, will predict emotions that occurred in a given segment.

This project aligns with Banijay Benelux's goal of gaining insights into viewer preferences and optimizing content creation strategies. By classifying emotions expressed in the series, Banijay Benelux can identify which segments resonate most with the audience, thereby informing future content decisions.

6.2 Analysis of Emotion Recognition Models

The understanding and interpretation of human emotions in textual data present both challenges and opportunities in various domains. By examining different datasets and employing various models, we gain insights into the complexities of emotional expression and the effectiveness of different approaches in capturing them.

Our analysis reveals that while simpler models like Naive Bayes offer efficiency, they may struggle with capturing the nuances of emotional expression due to their reliance on bag-of-words approaches. On the other hand, more advanced models like RNNs and CNNs show promise in capturing contextual dependencies and local patterns but may still face challenges in handling complex linguistic structures.

The Transformer models, particularly BERT, represents a significant advancement in emotion recognition. BERT's ability to leverage contextual information and linguistic nuances enables it to achieve superior performance across various metrics. However, its computational complexity and resource requirements may pose challenges for deployment in real-world applications.

6.3 Consideration of Dataset Annotation and Reliability

Furthermore, our exploration of different datasets highlights the importance of considering annotation methods and reliability. Manual annotation ensures high reliability and accuracy but may require significant time and resources. Distant supervision offers scalability but may introduce noise and inaccuracies, requiring additional preprocessing and quality control measures.

Overall, our findings underscore the importance of a comprehensive understanding of emotion recognition models, datasets, and evaluation metrics. By identifying strengths and limitations, we can inform the selection of appropriate models and methodologies for specific applications, ultimately advancing our understanding of human emotional expression in textual data.

6.4 Episode to Emotions Pipeline

As mentioned earlier, the pipeline we have created can predict emotions on a segment level from an entire episode in video format, and a table that divides the episode into segments using seconds from episode start. Our pipeline then cuts the video into smaller fragments, and extracts text automatically translated to English language using the OpenAI's Whisper state of the art model. This extracted text is then split into sentences, and pre-processed to be compatible with our best BERT transformer model. The model gives a prediction for each sentence, and checks if the most probable emotion reaches the threshold. If it does, the emotion is assigned to the sentence, and if the probability is too low, the sentence is labeled as neutral. Finally, unique predicted emotions are returned.

7 Limitations and Further Improvements

Due to the project lasting eight weeks, there are parts of the solution we would suggest to improve. The quality of training data is crucial in this task, and we believe that spending more time to find more reliable dataset or even creating a new one by hand would greatly increase the quality of the models used. On top of that, experimenting with other transformer architectures like RoBERTa might also be beneficial. Lastly, our current pipeline uses large whisper model to transcribe and translate audio files to text. While we believe this to be a solid solution, it may also be worth to check how this compares to first transcribing and later translating in a separate step. Additionally, looking for alternative technologies to both transcribe or translate could give better results.

8 Conclusion

Our examination of emotion recognition models and datasets provides valuable insights into the complexities of human emotional expression in textual data. Through thorough analysis and evaluation, we have identified the strengths and limitations of different models, ranging from Naive Bayes to advanced Transformer architectures like BERT.

While each model offers unique advantages and challenges, BERT emerges as a front-runner in terms of performance, leveraging its ability to capture contextual nuances and linguistic subtleties. However, the choice of model ultimately depends on the specific requirements of the application, considering factors such as computational resources, performance metrics, and dataset characteristics.

Our exploration of the client's needs, exemplified by Banijay Benelux's desire to gain insights into viewer engagement with TV shows, underscores the practical applications of emotion recognition technology. By developing tools for content classification and emotion tagging, we empower clients like Banijay Benelux to make informed decisions about their programming strategies and audience engagement initiatives.

Furthermore, our analysis of different datasets highlights the importance of annotation methods and reliability in ensuring the validity and generalizability of results. By understanding these factors and employing appropriate methodologies, we can enhance our understanding of human emotional expression and its computational interpretation.

Moving forward, our findings provide a foundation for further research and development in emotion recognition technology. Future efforts could focus on refining existing models, exploring hybrid approaches, and integrating user feedback for enhanced accuracy and applicability.

Our objective is to advance emotion recognition technology within the context of TV show analysis, particularly in collaboration with Banijay Benelux. By sharing our insights and recommendations, we aim to contribute to the improvement of emotion classification in this specific domain.

REFERENCES

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714.

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*.

APPENDIX 1