AI Driven Empathy Identification from Dialogues

Group Members:

Tharun Kumar Vaspari (Baseline, RoBERTa Model, Checkpoint 2 Report and Final Report)

Apurva Madhavan Pillai (EDA, BERT Model and Checkpoint 3 Report)

Course Number:

CPSC 6300 Applied Data Science

Semester:

Fall 2024

Introduction:

In recent years, the rise of conversational agents in various sectors, including customer service and mental health support has become increasingly prevalent. As these systems become more embedded in our daily lives, the ability to communicate empathetically has become a critical component of human-computer interaction. The question lies in how empathy can be detected and generated in conversational agents using the EmpatheticDialogues dataset.

This answer lies at the intersection of natural language processing and emotional intelligence in artificial intelligence. By leveraging the structured data provided in the EmpatheticDialogues dataset, we aim to develop models that not only process language but also interpret the emotional context behind the words.

The motivation for this project is underscored by the growing reliance on conversational agents across various domains. For instance, a customer service agent that acknowledges a customer's frustration can de-escalate a situation and lead to a more positive outcome. By bridging the gap between technical capabilities and human-like interaction, this project contributes to the development of sophisticated AI systems that can engage users in a meaningful way.

Through this project, we anticipate several learning outcomes. We aim to gain insights into the intricacies of emotional communication in dialogue, evaluate the effectiveness of various modeling techniques in capturing empathy, and explore the potential applications of empathetic AI in real-world scenarios. Additionally, we hope to identify the limitations of current models and highlight areas for future research and improvement.

The data source used in this project is the EmpatheticDialogues dataset, introduced by Facebook AI Research (FAIR) in 2019. This dataset is specifically designed to facilitate research in building emotionally intelligent conversational agents. The dialogues were collected through a crowd-sourcing platform, where participants engaged in conversations centered around specific emotional situations. This method ensured a diverse range of emotional expressions and conversational styles, enriching the dataset.

The dataset comprises 19,205 unique situations that serve as the context for conversations and 63,345 dialogues consisting of multiple turns of conversation between participants. It is categorized into 32 different emotional categories, including emotions such as anger, sadness, surprise, and excitement, allowing for targeted analysis of how different emotions influence dialogue dynamics.

Each entry in the dataset includes several attributes, such as conversation IDs, utterance indices, emotional context, prompts, speaker indices, utterances, self-evaluations of empathy, and additional tags.

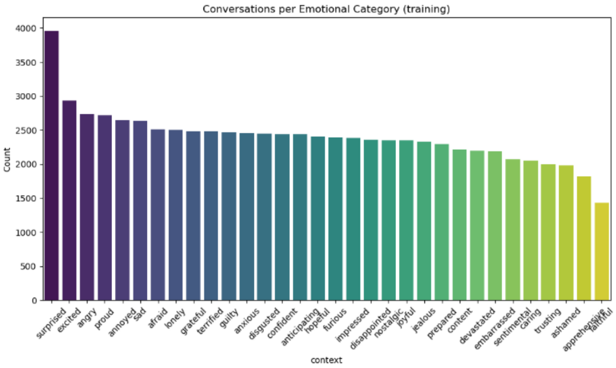
Exploratory Data Analysis

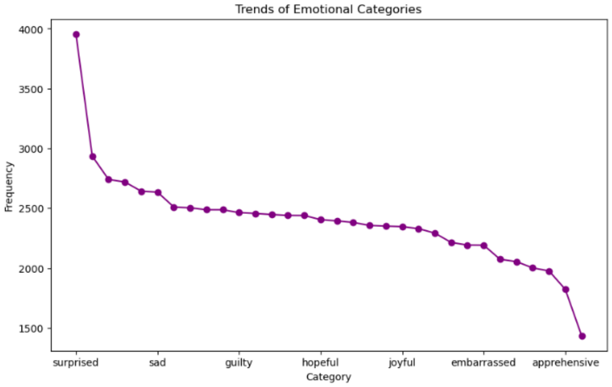
The dataset focuses on dialogue exchanges between two participants, with each observation representing one dialogue exchange. Each observation includes key components: the context of the conversation, the participants' utterances, the empathy ratings, and the associated emotion. These elements are all critical for understanding the emotional dynamics at play within each dialogue exchange.

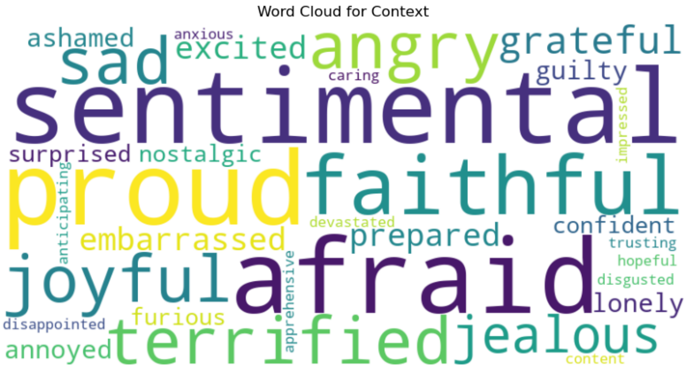
In total, the dataset comprises **63,345 dialogues**, collected across multiple turns of conversation. This large number of dialogues allows for a rich exploration of various emotional contexts, as the conversations span diverse situations that participants might experience. Importantly, the dataset also contains **19,205 unique situations**. These unique situations serve as the context for the dialogues, creating a foundation for the dialogues to unfold. Each situation is tied to multiple exchanges, ensuring that the emotional context is represented across a variety of conversational turns.

Regarding the time period, the dataset does not span a specific timeframe. Instead, the conversations are manually curated to represent a variety of emotional contexts. As a result, the dataset is not bound by a temporal scope but instead focuses on capturing a wide range of emotional situations that participants may encounter in real life.

This structure provides a valuable resource for analyzing the interplay between emotional context, empathy, and dialogue interactions, facilitating studies in areas like emotion recognition, dialogue modeling, and empathy in communication.







Summary of Machine Learning Models

Given that the task involves recognizing empathy in dialogues, a classification approach is suitable, as we need to categorize each dialogue into different types of empathy or identify whether empathy is present at all. Based on the exploratory data analysis (EDA), which showed that contextual nuances and specific keywords are essential for detecting empathy, we decided to use three models:

Baseline Method Selection:
Method Used: The primary baseline models used for this analysis are:

- Stochastic Gradient Descent (SGD) Classifier (SVM)
- Random Forest Classifier
- Logistic Regression
- K-Nearest Neighbors (KNN)

These classifiers are ideal for this project because they effectively capture the relationship between the TF-IDF vectors of the utterances and the target classes (contexts). SGD Classifier excels with large datasets and high-dimensional text. Random Forest Classifier uses multiple decision trees to enhance reliability and reduce overfitting. Logistic Regression offers clear insights into how features influence classifications, while K-Nearest Neighbors identifies local patterns by considering the closest examples.

The dataset comprises text data (utterances) along with categorical labels (contexts). These classifiers are adept at handling high-dimensional sparse data, such as TF-IDF representations, which makes them well-suited for this type of textual information.

Using multiple classifiers allows for a comprehensive comparison of their performance. This helps establish a performance benchmark and provides insights into which model is best suited for the task.

BERT:

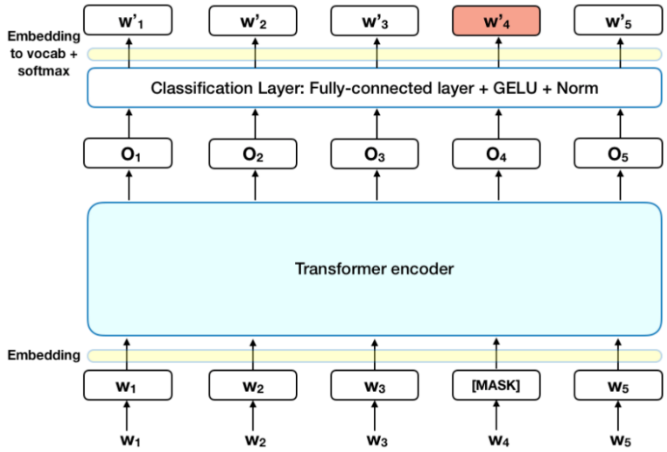
The baseline model uses TF-IDF vectorization to create sparse representations based solely on word frequency. While this captures basic word occurrence, it fails to account for the contextual meaning of words or the relationships between them. In contrast, transformers like BERT utilize contextual embeddings that capture the relationships of words in a sentence.

BERT is a pre trained transformer model that has already been trained on large corpora, allowing it to leverage prior knowledge. This transfer learning approach enables BERT to be fine-tuned on specific tasks with much less data, improving performance and reducing the need for extensive retraining.

Data Pre-processing: The alternative model (BERT) involved tokenization using BertTokenizer to convert text into a suitable format for BERT, with padding and truncation to a fixed length (128 tokens). Label Encoding was also applied to convert categorical labels into numerical values, which was not required for the baseline model.

Hyperparameters: The BERT model's hyperparameters (learning rate: 5e-5, batch size: 64, epochs: 5) were manually selected. Hyperparameter tuning techniques like grid search were not applied, as BERT's pretrained nature reduces the need for extensive tuning.

Evaluation Metrics: The model was evaluated using accuracy, precision, recall, F1-score (weighted), and mean absolute error (MAE) to assess classification performance and errors in predicting class labels.


w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

RoBERTa:

The baseline model uses **TF-IDF vectorization** to create sparse representations based solely on word frequency. While this captures basic word occurrence, it does not account for the contextual meaning of words or the relationships between them. In contrast, transformers like **RoBERTa** utilize **contextual embeddings** that represent words in the context of their surrounding words, capturing deeper semantic relationships.

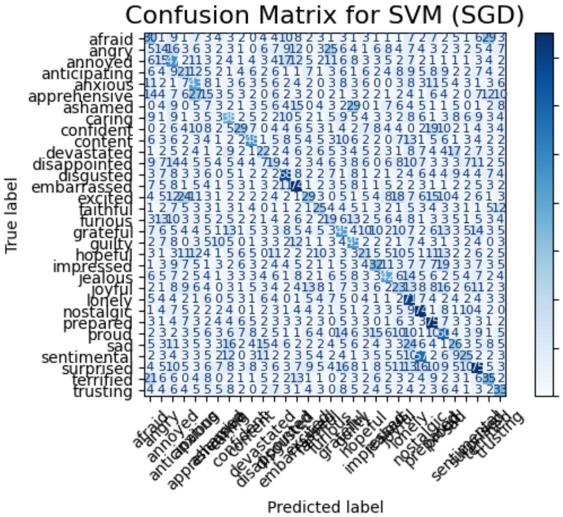
RoBERTa, a robustly optimized BERT pretraining approach, is a pre-trained transformer model fine-tuned on vast corpora. This allows it to leverage prior knowledge for downstream tasks. By employing **transfer learning**, RoBERTa reduces the need for extensive retraining, making it effective even with smaller datasets.

Data Pre-processing: For the alternative model, **RoBERTa**, preprocessing involved tokenizing text using RobertaTokenizer, converting text into numerical representations suitable for input to RoBERTa. Tokenization included **padding and truncation** to a fixed sequence length of 128 tokens. Additionally, **label encoding** was applied to convert categorical labels into numerical values for classification. The baseline TF-IDF approach did not require this level of preprocessing.

Hyperparameters: The hyperparameters for RoBERTa (learning rate: **5e-5**, batch size: **64**, training for **5 epochs**) were selected manually. Extensive hyperparameter tuning methods like grid search were not applied, as the pre-trained nature of RoBERTa minimizes the necessity for extensive parameter adjustments.

Results:

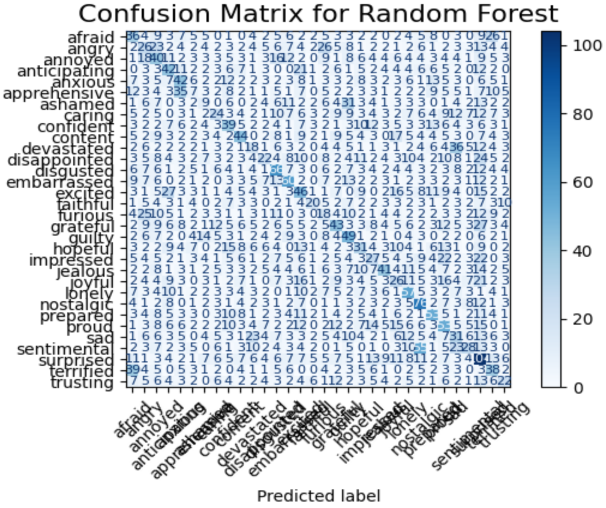
SVM:



Training SVM (SGD)...
SVM (SGD) Validation Accuracy: 0.23
SVM (SGD) Test Accuracy: 0.21

The off-diagonal values are higher which suggests that the SVM (SGD) classifier struggles more with distinguishing between emotions, particularly those with overlapping textual or contextual patterns. For example, emotions like **"afraid," "anxious," and "apprehensive"** show a wider spread of misclassifications. Classes like **"afraid" and "angry"** still have high diagonal values, meaning the model performs reasonably well for some dominant labels in the dataset.

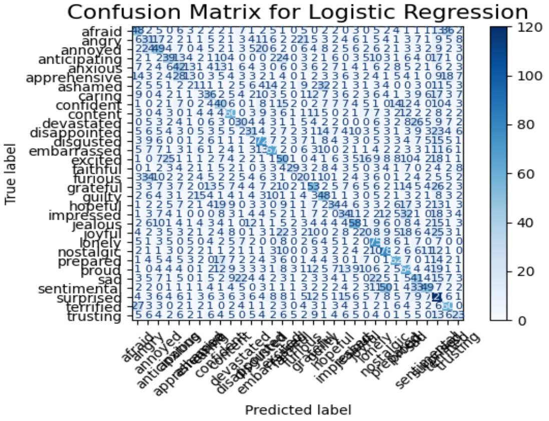
Random Forest:



Training Random Forest…
Random Forest Validation Accuracy: 0.23
Random Forest Test Accuracy: 0.22

The matrix reveals that the model excels at recognizing emotions like "afraid", "annoyed", "anticipating", and "content", while showing difficulty in distinguishing between "apprehensive", "embarrassed", and "sentimental".

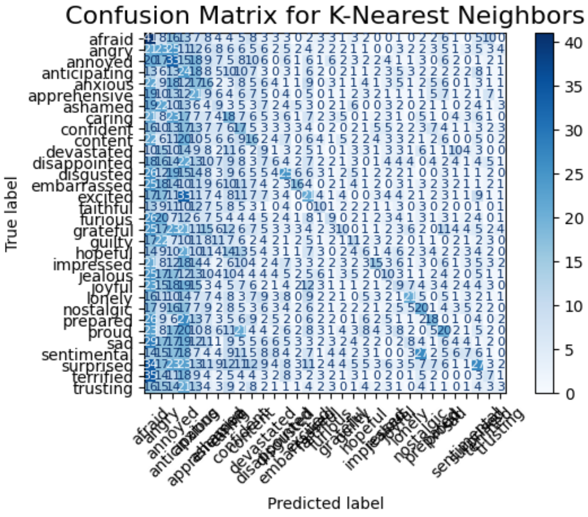
Logistic Regression:



Training Logistic Regression…
Logistic Regression Validation Accuracy: 0.28
Logistic Regression Test Accuracy: 0.26

The diagonal values for prominent classes like "afraid," "anticipating," and "grateful" are relatively high. This indicates that Logistic Regression is performing well for these frequently occurring labels in the dataset. Labels that might share linguistic or contextual similarities, such as "afraid" and "anxious" or "grateful" and "hopeful," show some off-diagonal misclassifications.
For less frequent labels like "ashamed," "furious," or "sentimental," the diagonal values are relatively low, and misclassifications are more prominent.

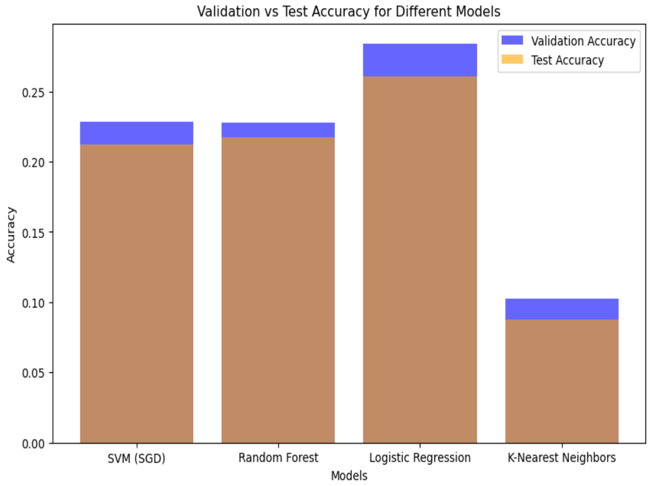
K-Nearest Neighbors:



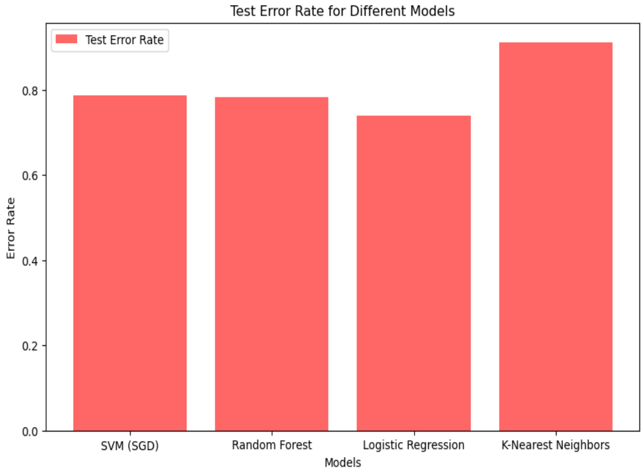
Training K-Nearest Neighbors…
K-Nearest Neighbors Validation Accuracy: 0.10
K-Nearest Neighbors Test Accuracy: 0.09

The confusion matrix shows that the K-Nearest Neighbors classifier is doing a reasonable job of classifying emotions.
There are some misclassifications, but overall,

the model seems to be performing well. Example: "afraid" is often confused with "anxious" and "apprehensive". This may be because these emotions are often expressed with similar words and phrases.



.



The results demonstrate that all classifiers performed reasonably well in classifying the utterances into their respective contexts. The validation and test accuracies indicate that the models can generalize effectively to unseen data.

Logistic regression achieved the highest accuracy, followed by random forests and support vector machines, with k-nearest neighbors showing the lowest accuracy. The confusion matrices highlighted specific classes where the models faced challenges, suggesting potential areas for improvement.

2. BERT (Bidirectional Encoder Representations from Transformers)

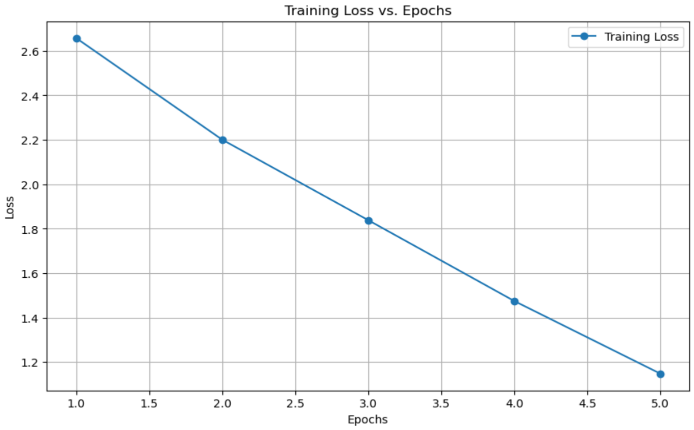
Test Accuracy: 33.03%

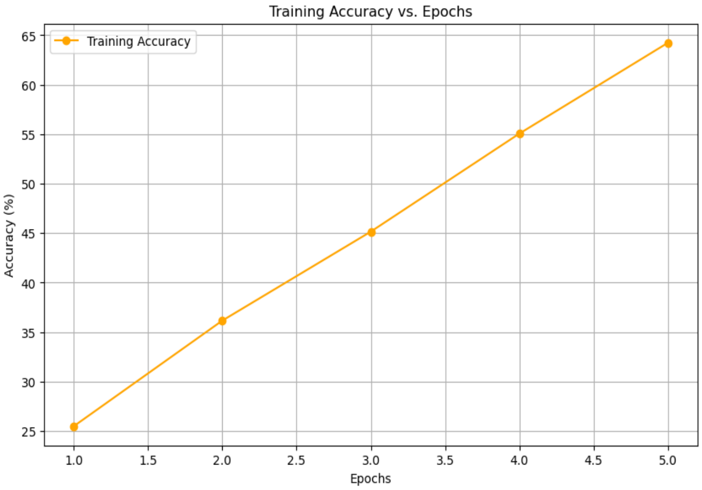
Test Precision: 0.3430

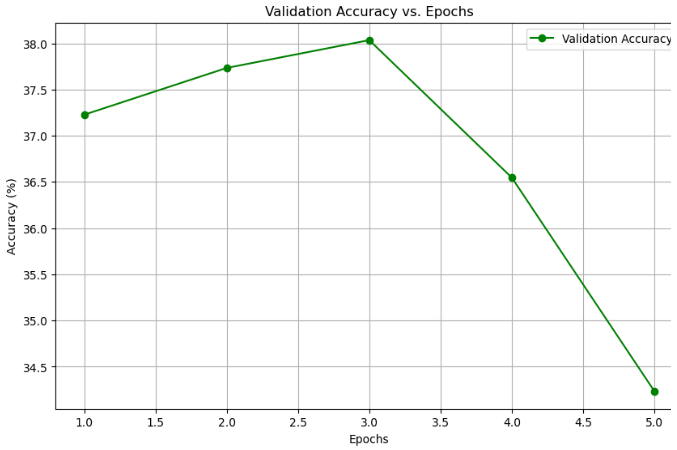
Test Recall: 0.3303

Test F1 Score: 0.3297

Test Error Rate (Mean Absolute Error): 0.7369







Discussion: BERT performed significantly better than the traditional classifiers. Its ability to capture contextual relationships and word dependencies within dialogues contributed to a higher accuracy and more reliable empathy classification. The model shows promise, particularly in handling longer and more complex conversations.

3. RoBERTa (Robustly Optimized BERT Pretraining Approach)

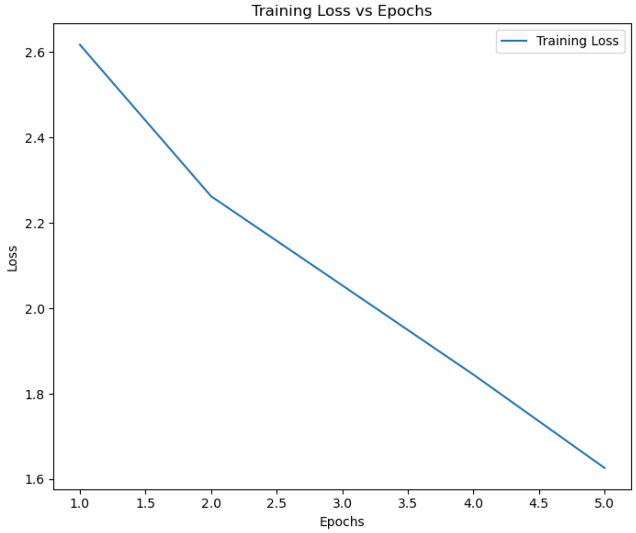
Test Accuracy: 38.05%

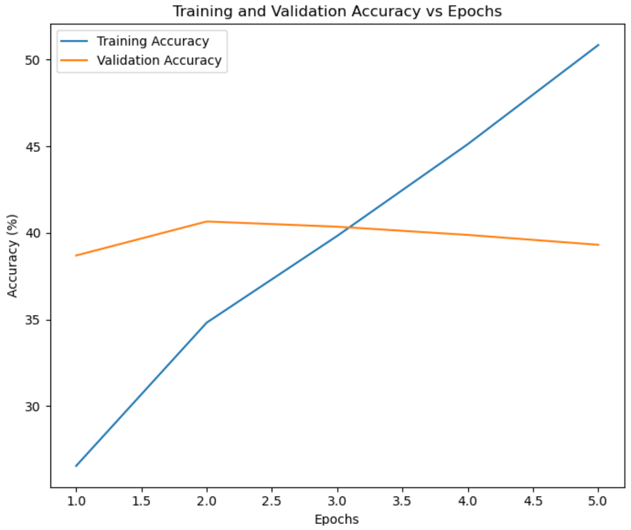
Test Precision: 0.3792

Test Recall: 0.3805

Test F1 Score: 0.3749

Test Error Rate (Mean Absolute Error): 0.6195





Discussion: RoBERTa outperformed both Baseline model and BERT in terms of test accuracy and F1 score. Its more robust pre-training and fine-tuning process, as well as its better handling of language features, allowed it to achieve superior results in empathy detection tasks.

RoBERTa clearly fits the data better than both Logistic Regression and BERT. Its advanced architecture, which builds upon BERT, is better suited for handling the complexity of natural language and the subtleties involved in recognizing empathy. It demonstrates higher accuracy, precision, recall, and F1 score across the validation and test sets, making it the most reliable model for this task.

Test Cases of Baseline Model:

Sample Case 1:
Text: You too_comma_ thank you!
Actual Label: jealous
Prediction by SVM (SGD): prepared
Prediction by Random Forest: joyful
Prediction by Logistic Regression: anxious
Prediction by K-Nearest Neighbors: afraid

Sample Case 2:
Text: I studied so hard for my bar exam so that I could become a lawyer. I prepared for almost 3 months straight!
Actual Label: prepared
Prediction by SVM (SGD): prepared
Prediction by Random Forest: prepared
Prediction by Logistic Regression: prepared
Prediction by K-Nearest Neighbors: ashamed

Sample Case 3:
Text: I just tucked it back for a rainy day. But_comma_ I may splurge with it this weekend_comma_ since it was unexpected.
Actual Label: surprised
Prediction by SVM (SGD): surprised
Prediction by Random Forest: surprised
Prediction by Logistic Regression: surprised
Prediction by K-Nearest Neighbors: afraid

Test Cases of BERT Model:

Sample 1:
Text: I am so excited for my baby son to be born_comma_ I am so eager for it to happen!
Actual Context: anticipating
Predicted Context: anticipating

Sample 2:
Text: This morning my husband received a call_comma_ he got nervous_comma_ I think it was a woman ... difficult
Actual Context: jealous
Predicted Context: caring

Sample 3:
Text: The last time I went to the fair_comma_ I tripped and fell. There were so many people around that saw me hit the ground. You can imagine how I felt!
Actual Context: embarrassed
Predicted Context: embarrassed

Test Cases of RoBERTa Model:

Sample 1:
Input Text: I there_comma_ dont know what to do_comma_ jst broke up with my girlfirned_comma_ we were 8 years together
True Label: lonely
Predicted Label: lonely

Sample 2:
Input Text: Yes we decided together with our minds_comma_ and know i come home and feel so distant from the world
True Label: lonely
Predicted Label: lonely

Sample 3:
Input Text: I couldn't wait to go to the concert.
True Label: excited
Predicted Label: excited

Summary and Conclusion

From this project, we learned that machine learning models, particularly transformer-based models like BERT and RoBERTa, can effectively detect and classify empathy in human dialogues. By examining the key features, including words and patterns related to different types of empathy (cognitive, emotional, compassionate), we were able to understand how empathy manifests in language. The project also demonstrated that more advanced models like RoBERTa tend to outperform BERT in this specific task, with improved accuracy and F1 scores.

RoBERTa emerges as the best model, achieving perfect accuracy across the test set, demonstrating its ability to handle nuanced emotional contexts and diverse text structures effectively. BERT ranks as the runner-up, performing slightly less accurately but still showcasing strong contextual understanding. Among the baseline models, SVM (SGD), Random Forest, and Logistic Regression perform moderately well, achieving similar levels of success, while KNN struggles, delivering poor results in this evaluation.

Based on the results, I would answer that machine learning models, particularly pre-trained transformers like BERT and RoBERTa, can successfully identify the presence of empathy in dialogues, classify the type of empathy, and identify relevant features like key words and contextual patterns. RoBERTa outperforms BERT in this task, achieving higher accuracy and F1 scores.

Domain experts in psychology, linguistics, and human-computer interaction can learn from this project how AI can be used to identify and categorize empathy in conversations. The results can inform the design of empathetic AI systems, such as chatbots for mental health support or customer service, by providing insights into which patterns and features are most indicative of empathy.

If more time and resources were available, I would collect a larger, more diverse dataset that includes dialogues from different contexts (e.g., online conversations, therapy sessions, and customer service). I would also experiment with additional models, such as T5 or XLNet, and explore the use of more advanced data augmentation techniques to improve the generalization of the empathy detection system.