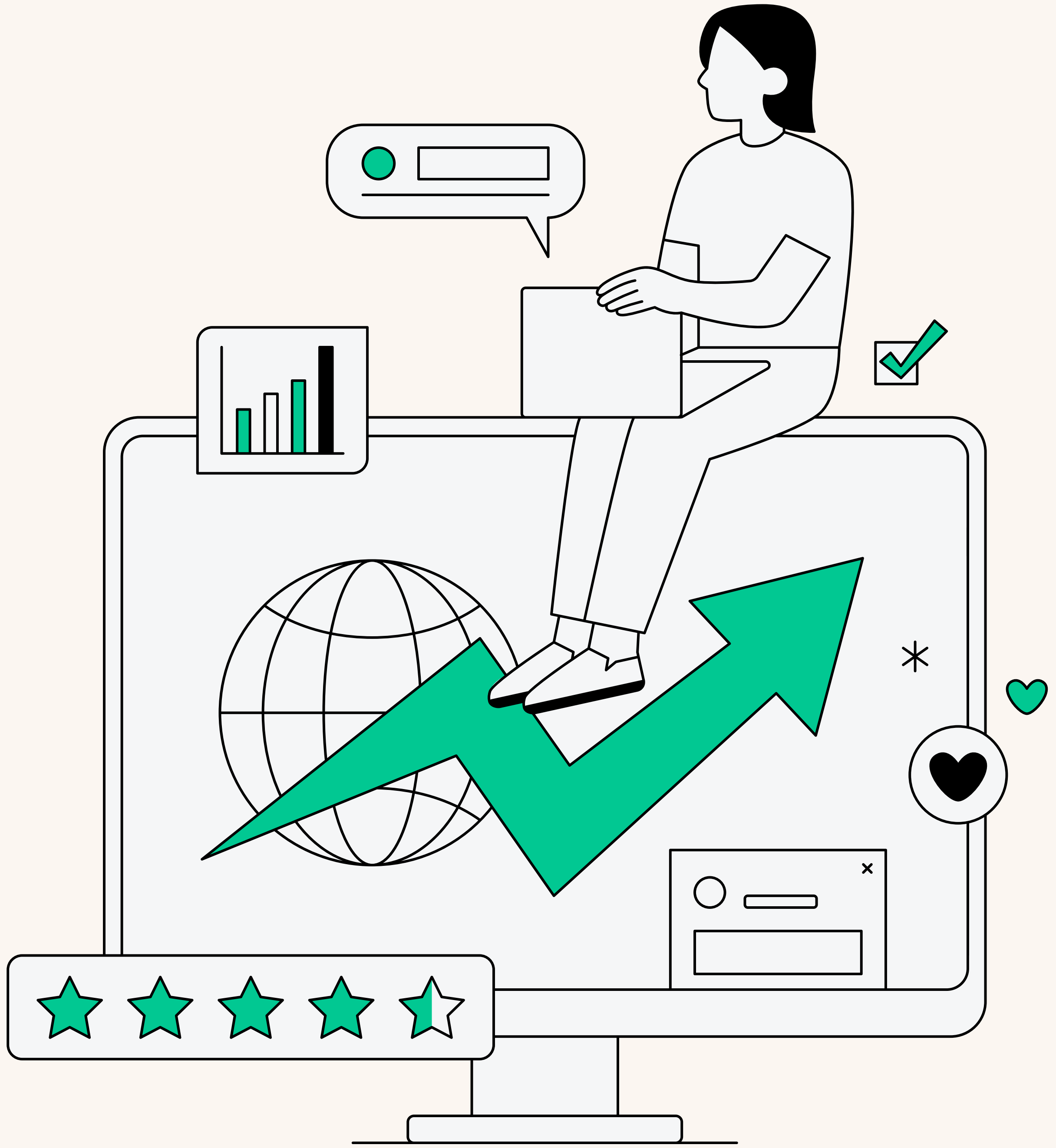
Presented by Mahfooz Anas (514587)
Arib Zayn Araf (514550)

CLUELESS

Contextual ***L***earning for ***U***nderstanding ***E***motion with ***L***imited ***E***xposure to ***S***emantic ***S***ymbols



2024-2025

What are our goals?

Explicit goals:

- Compare the performance of pre-trained models RoBERTa and EmotionBERT fine-tuned on the GoEmotions dataset.

- Implement and evaluate a masking strategy (we will see it later)

- Identify the best model and hyperparameters through grid search and analysis of metrics like Macro F1 Score.

Implicit goals:

- Apply key NLP techniques, such as tokenization, masking, and fine-tuning pre-trained models.
- Gain practical experience in handling and preprocessing real-world text datasets, including masking emotional keywords and creating expanded emotion word lists.
- Develop hands-on skills in optimizing hyperparameters and evaluating models with metrics beyond accuracy, such as Macro F1 and Weighted F1 Scores.

Why “CLUELESS”?

The project is named "CLUELESS" to emphasize the masking strategy applied during preprocessing. Words that directly signify emotions were masked 10% of the time. This was done to:

- ***Encourage Contextual Learning:*** The model couldn't rely on explicit emotional words as shortcuts but instead had to infer emotions based on the context of the surrounding text.
- **Improve Generalization**: By avoiding dependence on specific words, the model became less biased and better at handling nuanced cases, such as sarcasm or indirect emotional expressions.

This innovative masking method challenges the model to perform effectively even when obvious semantic "clues" are hidden, making it truly "clueless."

Possible Usage of This Method:

The masking strategy used in this project has broader applications beyond emotion classification, including:

Mitigating Racial Bias:

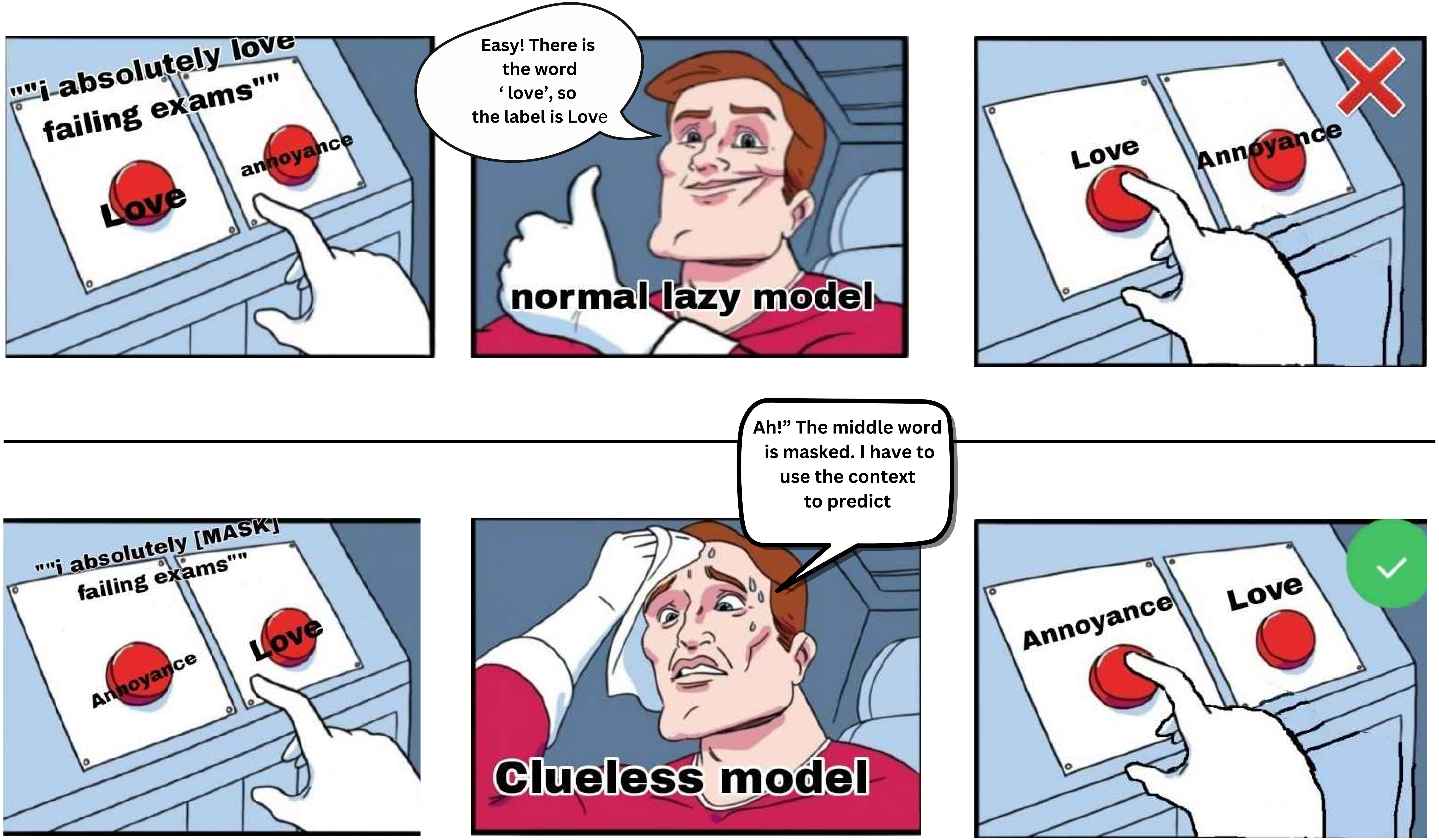
- By masking racial slurs or words tied to specific ethnic groups, the method prevents models from associating these words with biased sentiments.
- For example, in datasets where certain ethnic terms are disproportionately linked to negative or positive emotions, masking ensures the model doesn't propagate harmful stereotypes.

Reducing Societal Bias:

- This technique can be applied to sensitive domains where biased language (e.g., gender-specific terms or slurs) frequently appears.
- It reduces risks of the model unintentionally reinforcing or amplifying societal stereotypes in tasks like recruitment or content moderation.
- This approach may fosters ethical AI by ensuring our NLP models are less prone to bias, improving fairness, inclusivity, and generalization in real-world applications.

Addressing Ambiguity in Sentiment Analysis:

- Masking can also handle challenges like sarcasm or indirect language by forcing the model to rely on context rather than specific explicit words, making it more robust in detecting nuanced sentiment.


""i absolutely love failing exams""
Love
annoyance
Easy! There is the word ‘ love’, so the label is Love
normal lazy model
Love
Annoyance
""i absolutely [MASK] failing exams""
Annoyance
Love
Ah!” The middle word is masked. I have to use the context to predict
Clueless model
Annoyance
Love

Emotion Bert vs RoBerta.

- We fine-tuned two pre-trained models, EmotionBERT and RoBERTa, on the GoEmotions dataset with 27 emotions. We masked 10% of the emotional words, forcing the models to rely on context, even in tricky cases like sarcasm.

- After running grid searches with 3 learning rates, 4e-5 came out as the champ. Both models trained on 90% of the dataset using AdamW optimizer, CrossEntropyLoss, and a 16-batch size for 2-epochs. We kept an eye on validation loss to avoid overfitting.

- In the end, we compared their performance using accuracy, macro F1, and weighted F1 scores to see which one was the real MVP under the masked setup.

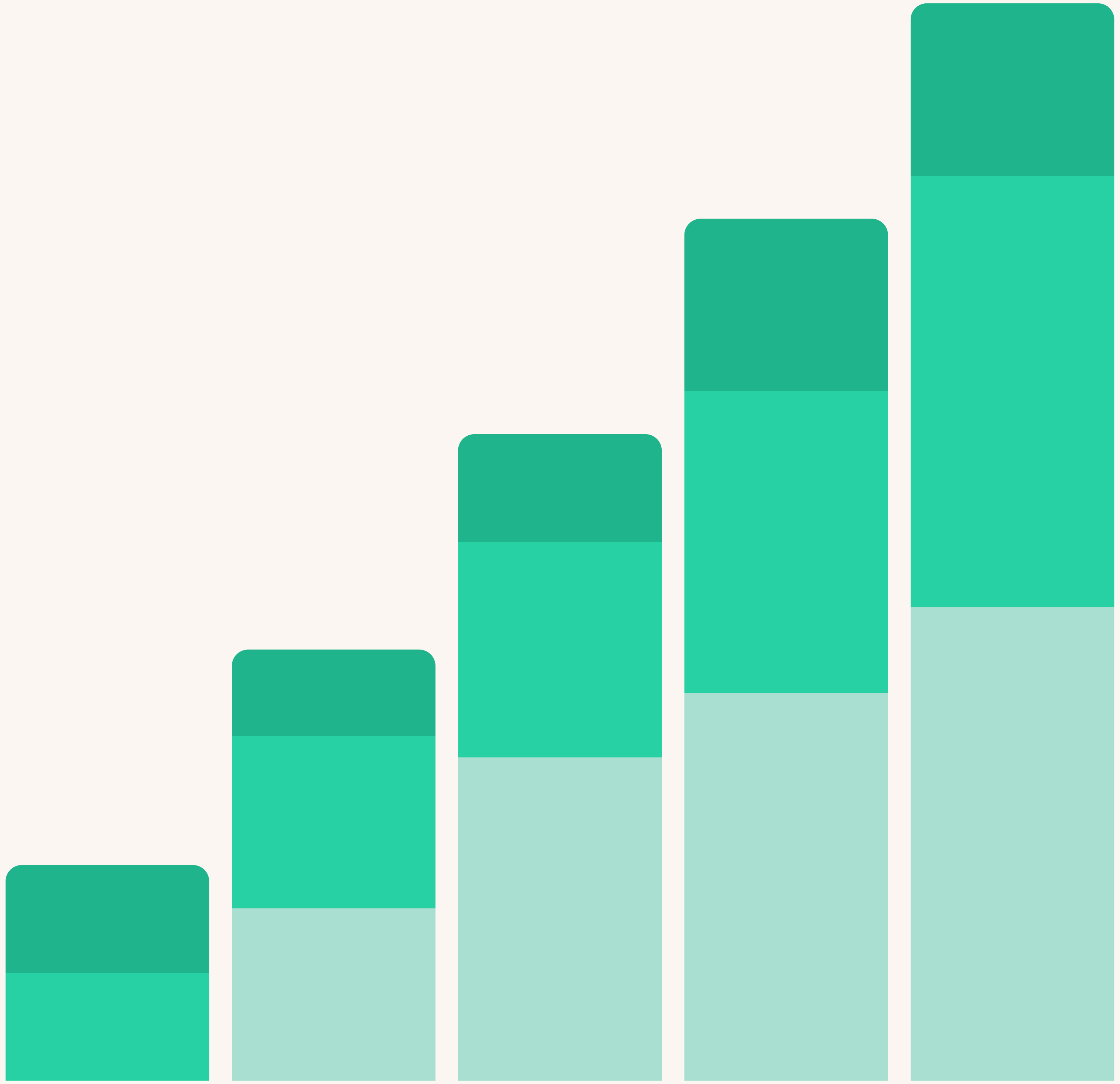
RoBERTa:

- RoBERTa (Robustly Optimized BERT Pretraining Approach) is an improved version of BERT, trained on more data (160GB) and with longer training time.

- Trained on datasets like BookCorpus, Common Crawl News, OpenWebText, and Stories

EmotionBERT:

- EmotionBERT is a variation of BERT fine-tuned specifically for emotion recognition in text.
- It is trained on emotion-labeled datasets, such as raw GoEmotions, SemEval, and ISEAR, to identify emotions like joy, sadness, anger, and fear.
- Used for sentiment analysis, chatbot interactions, and affective computing.



Expectations

CHRISTIANO RONALDO WINNNIG THE 2022 FOOTBALL WORLD CUP.

- EmotionBERT was pre-trained on the raw GoEmotions dataset, so we expected it to perform better as it is already familiar with the dataset's structure and emotional labels.

- RoBERTa, being a general-purpose model, was expected to be less accurate as it wasn't pre-trained for emotion-specific tasks.

- The masking strategy was anticipated to improve generalization for both models, but EmotionBERT was likely to adapt better due to its task-specific pretraining..

- We Expected EmotionBERT's specialized training would give it an edge over RoBERTa in terms of accuracy and F1 scores.

Reality !!!

LIONEL MESSI WON THE 2022 FOOTBALL WORLD CUP.

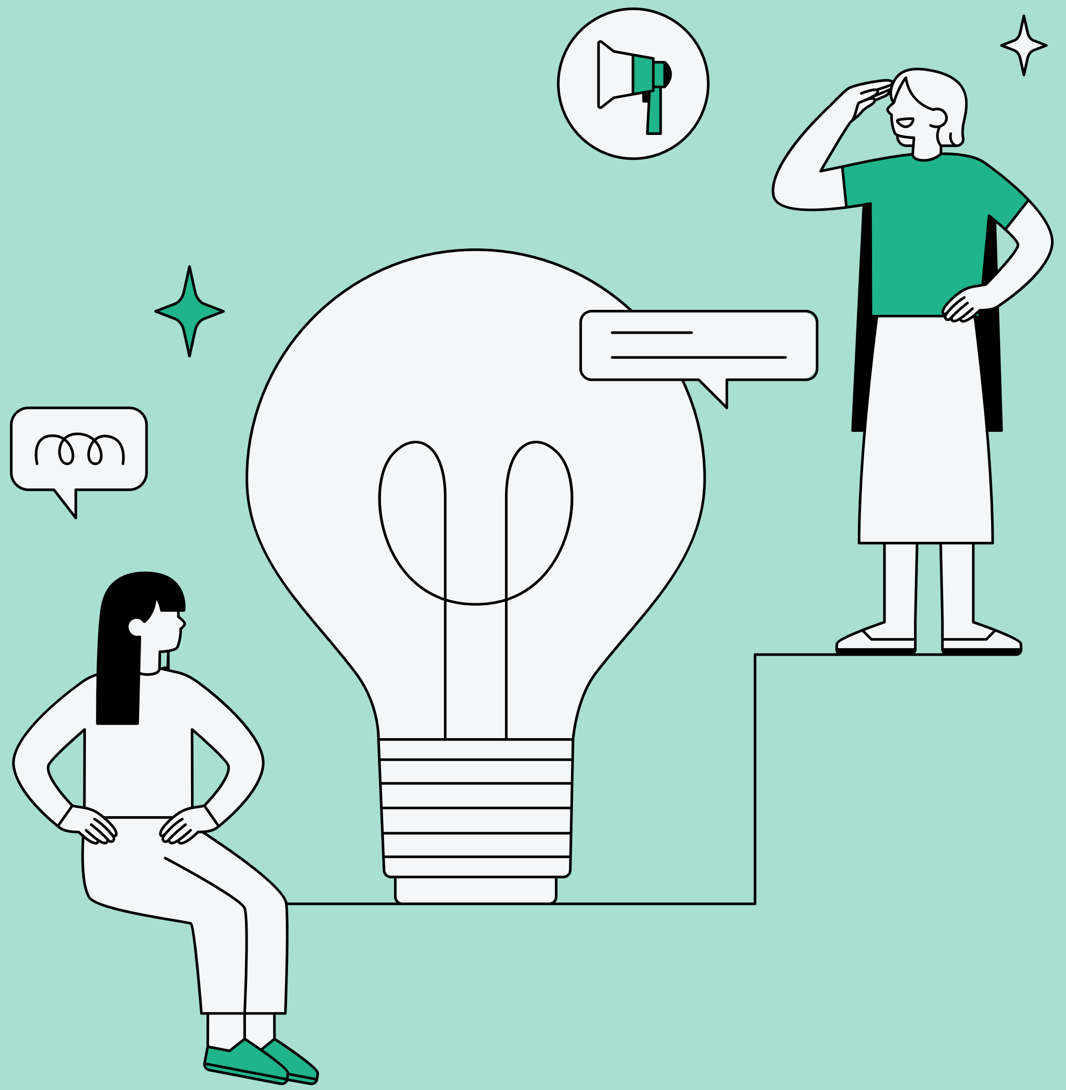
RoBERTa slightly outperformed EmotionBERT, though the difference was smaller, despite one of the dataset it was pre trained on was raw Goemotion dataset.

	RoBERTa	EmotionBERT
Accuracy	0.57	0.52
Weighted F1 score	0.42	0.40
Macro F1 score	0.55	0.51

Performance Metrics

- Both models adapted to the masking strategy, which would force reliance on context rather than explicit emotional keywords.
- The macro f1 score showing that RoBERTa classified labels correctly for every single labels, better than EmotionBERT, although the score is pretty close.

The Question Remains



Did masking actually work?

It has been assumed that masking emotion specific words will help generalization by context better, because of the nature of pre-trained models used, but it wasn't evaluated extrinsically.

Possible solution: Apply the same procedure,with the same hyperparameters, but this time without masking strategy, and compare the f1 scores.

Did roberta really do better or is it just better at handling masked words?

Possible solution : same as the first one, but compare f1 scores of RoBERTa model and EmotionBERT, without masking strategy.

Was the choice of hyperparameter optimal?

The grid search used on learning rate with smaller portion of dataset to find the best learning rate, assumes that the this will remain the best learning rate even with bigger dataset, which may not be the case.

Possible solution : apply optuna bayesian optimization on the whole datatset to find the best hyperparameters. Although it will be resource costly.

Presented by

Anas Mahfooz (514587)
Araf Arib Zayn (514550)

Thank you very much!

2024-2025