Emotion Discovery and Reasoning its Flip in Conversation

NLP Course Project

Alessandro Pasi
Matteo Belletti
Stricescu Razvan Ciprian

Introduction

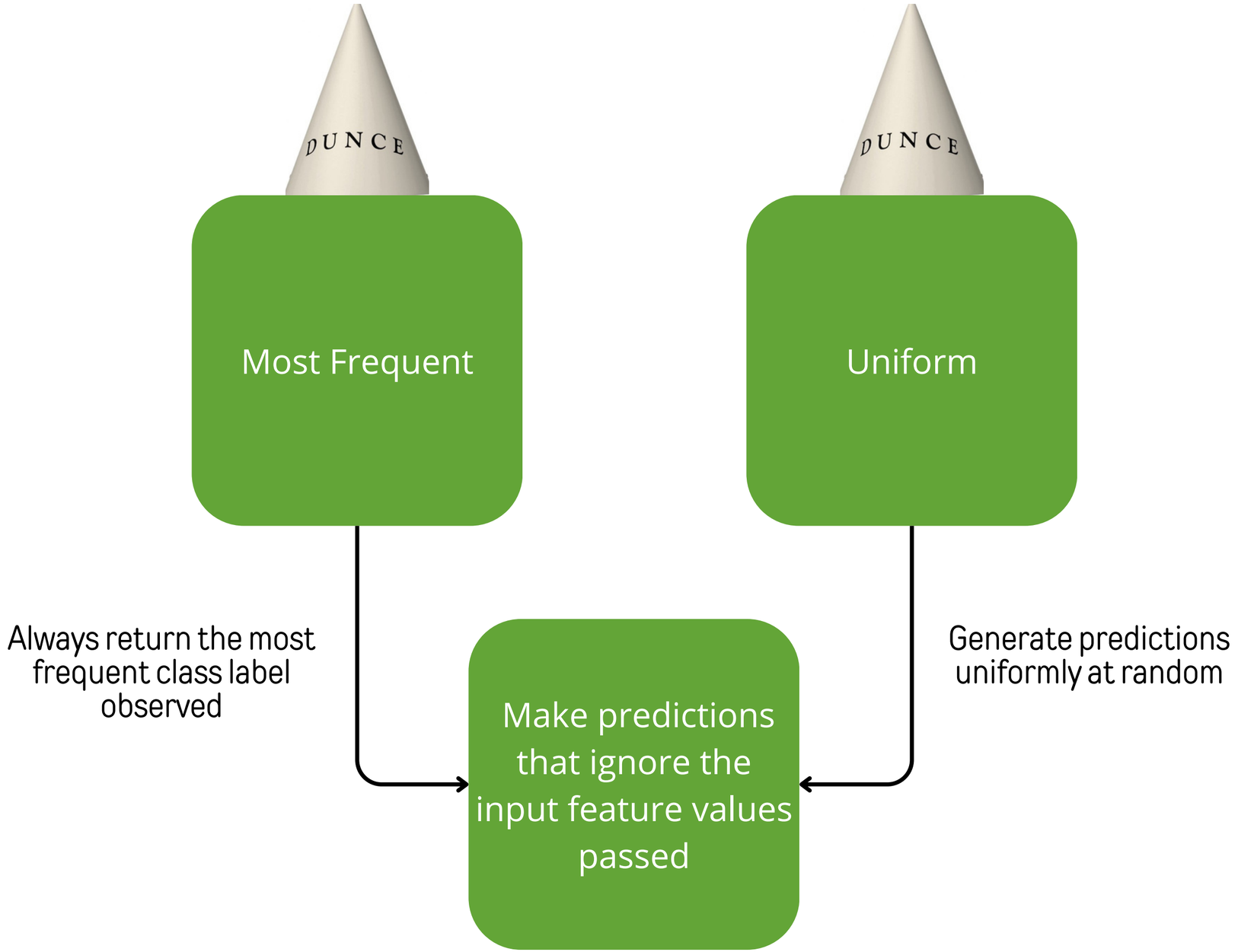
- We aim to detect emotions and identify emotional shifts in multi-party conversations through various model architectures and techniques. Our comprehensive approach involves creating multiple models, including BERT-based architectures inspired by recent research.

The dataset used is based upon dialogues from the TV show Friends.

	True labels	
	Emotions	Triggers
S1: Hey, so uh, y'know how there's something I wanted to talk to you about?	Neutral	0
S2: Oh yeah!	Joy	0
S1: Well, y'know how I'm trying to work things out with Emily.	Joy	0
S2: Well, there's this one thing… Okay, here goes.	Neutral	0
S1: I made a promise that--Oh hey!	Surprise	0
S2: What?	Neutral	0

System description

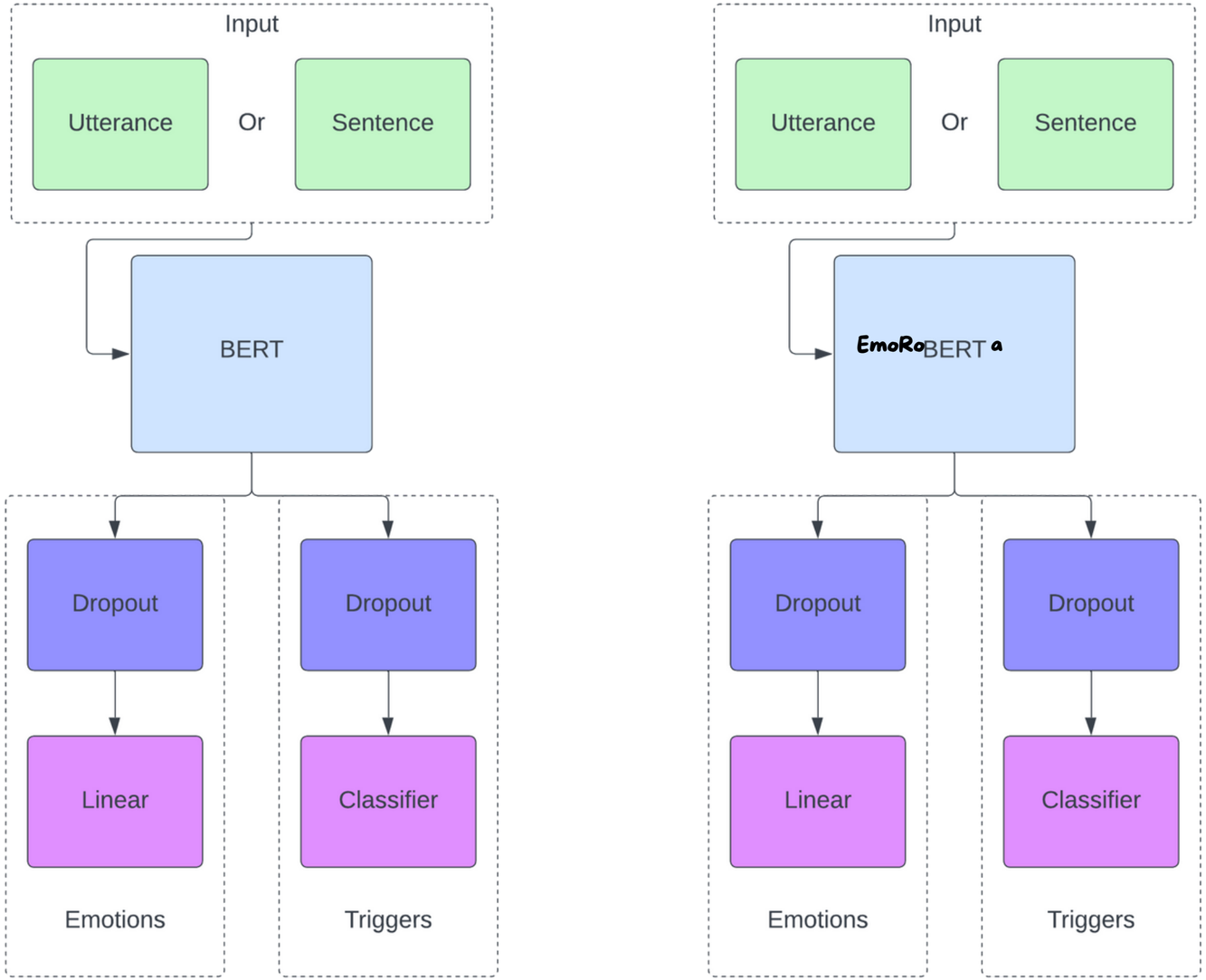
Initially, we established baseline performance with two dummy models using "uniform" and "most frequent" strategies via DummyClassifier from the sklearn library. These served as comparisons for our personalized model based on BERT and EmoRoBERTa.



System description

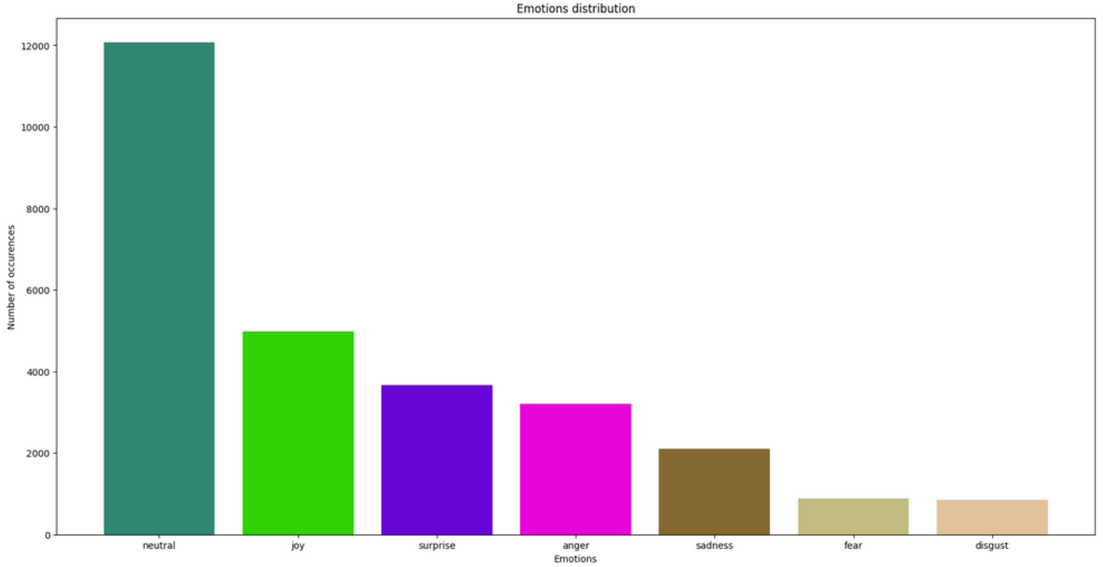
The BERT models share a standardized architecture. We augmented this architecture with two personalized classification heads for 'trigger' and 'emotion' labels, each comprising a dropout layer and a linear layer with outputs tailored for emotions and triggers.

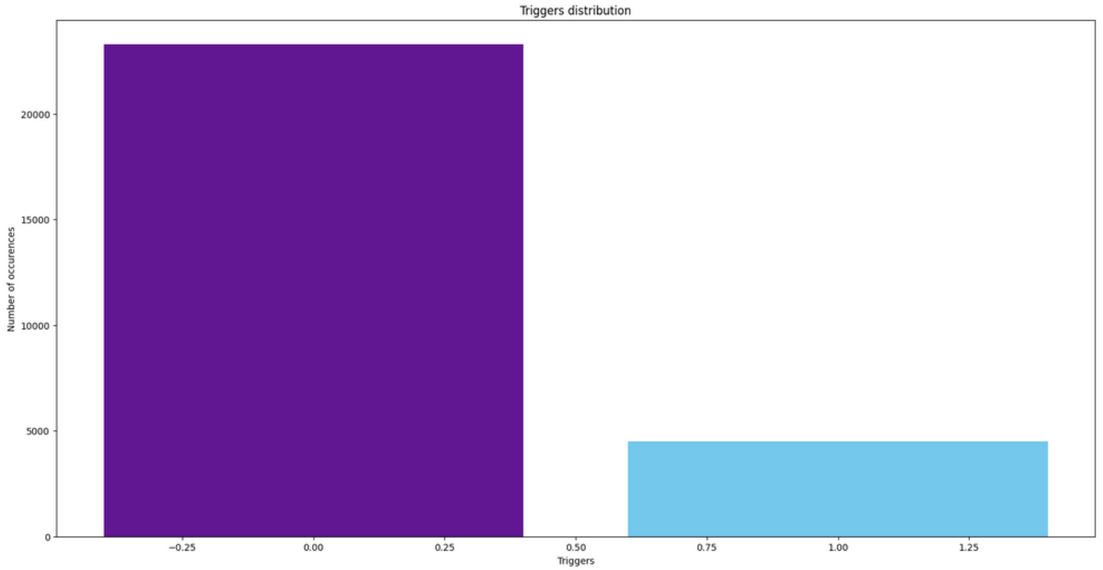
Lastly, we introduced EmoRoBERTa embeddings to assess their potential to enhance emotion recognition performance.



Data

Visualizing the data distribution revealed significant class imbalances, particularly towards neutral emotions and non-trigger instances. To mitigate this, we adopted weight balancing to prioritize underrepresented labels during model training.





Context analysis

- Initially, for the BERT models, we attempted a teacher-forcing technique by predicting labels for entire dialogues based on sentence-by-sentence input. However, this approach yielded suboptimal results, possibly due to short dialogues rendering context less informative and introducing noise. Subsequently, we simplified the approach to single-sentence input, which significantly improved performance compared to the contextual input method. While intuitive for emotion labels, this approach was unexpectedly effective for trigger prediction despite its presumed context-dependence.

Experimental setup and results

- Computational Infrastructure: Our experiments were conducted on both Colab and local GPUs.
- Baseline classifiers: We implemented a random and a majority classifier
- Common Architecture: All proposed BERT models share identical architectures.
- Initial Testing: For initial Base Bert testing we utilized a smaller version of it.
- Model Evolution:
  - First model: Single-head classification layer on "context" dataset.
  - Second model: Two similar classification heads, each with dropout and linear layers.
- Dataset Simplification: Transitioned from "context" dataset to individual sentence inputs, resulting in improved model performance.
- Frozen vs. Unfrozen Models: Implemented both versions to assess the impact of embedding layers.

Experimental setup and results

- Hyperparameter Optimization:
  - Batch size reduction to handle computational constraints.
  - Adopted hyperparameters proposed in (Devlin et al., 2019) for efficiency.
- Optimizer and Scheduler Choice: Utilized Adam optimizer and StepLR scheduler for simplicity and effectiveness.
- Custom Loss Function: Developed a custom loss function to handle the dual nature of the classification task.
- Class Weighting: Introduced class weights to counter dataset class imbalance, resulting in improved performance.
- Changing the embeddigs: We implemented a EmoRoBERTa model using the same architrecture as the preveous BERT models
- Evaluation Metrics:
  - Focused primarily on macro and weighted F1-scores.
  - Evaluated models based on unrolled sequences and dialogue-based aspects.

Results

Our experiment compared two models: one trained with context and one without. Surprisingly, the non-contextual model consistently outperformed the contextual one in F1-scores, prompting questions about context's importance. Further analysis showed only a small percentage of emotions and triggers depend on context, supporting the simpler model's credibility.

Seed		4	42
Model with "context"			
Unrolled	Macro	E: 0.15 T: 0.46	E: 0.17 T: 0.46
	Weighted	E: 0.31 T: 0.79	E: 0.30 T: 0.79
Sequence	Dialogue average	E: 0.16 T: 0.45	E: 0.14 T: 0.45
Model without "context"			
Unrolled	Macro	E: 0.20 T: 0.46	E: 0.32 T: 0.49
	Weighted	E: 0.38 T: 0.79	E: 0.47 T: 0.78
Sequence	Dialogue average	E: 0.31 T: 0.45	E: 0.34 T: 0.46
EmoRoBERTa w "context"			
Unrolled	Macro	E: 0.04 T: 0.46	E: 0.04 T: 0.46
	Weighted	E: 0.05 T: 0.79	E: 0.05 T: 0.79
Sequence	Dialogue average	E: 0.09 T: 0.45	E: 0.09 T: 0.45
EmoRoBERTa w/o "context"			
Unrolled	Macro	E: 0.26 T: 0.46	E: 0.31 T: 0.46
	Weighted	E: 0.36 T: 0.79	E: 0.48 T: 0.79
Sequence	Dialogue average	E: 0.32 T: 0.45	E: 0.37 T: 0.45

Seed	4		42	
RMSE	U	S	U	S
Bert w "context"	E: 2.12 T: 0.38	E: 0.09 T: 0.04	E: 2.27 T: 0.38	E: 0.09 T: 0.04
Bert w/o "context"	E: 1.99 T: 0.38	E: 0.16 T: 0.04	E: 1.94 T: 0.41	E: 0.16 T: 0.06
EmoRoBERTa w "context"	E: 1.80 T: 0.38	E: 0.07 T: 0.04	E: 1.80 T: 0.38	E: 0.07 T: 0.04
EmoRoBERTa w/o "context"	E: 1.80 T: 0.38	E: 0.18 T: 0.04	E: 1.90 T: 0.38	E: 0.19 T: 0.04s

Results

In conclusion, the unfrozen configuration proved more effective for BERT-based approaches, allowing the models to learn embeddings better. However, contrary to expectations, the frozen embedding layer performed better for EmoRoBERTa models. This is notable because EmoRoBERTa models already possess highly effective emotion embeddings, and further training led to a decline in performance. This underscores the counterproductivity of unnecessary adjustments in such scenarios.

Model	Emotion		Trigger	
Macro F1	Mean	S.d.	Mean	S.d.
Majority	0.8	0	0.46	0
Random	0.126	0.005	0.428	0.008
BERT context	0.11	0.064	0.46	0
BERT context frozen	0.08	0.066	0.46	0
BERT w/c	0.234	0.097	0.468	0.013
BERT w/c frozen	0.198	0.004	0.46	0
EmoRoBERTa	0.04	0	0.46	0
EmoRoBERTa frozen	0.04	0	0.46	0
EmoRoBERTa w/c	0.254	0.098	0.462	0.004
EmoRoBERTa w/c frozen	0.288	0.03	0.46	0

Model	Emotion		Trigger	
Macro F1	Mean	S.d.	Mean	S.d.
Majority	0.08	0	0.46	0
Random	0.126	0.005	0.428	0.008
BERT	0.124	0.036	0.45	0
BERT frozen	0.124	0.036	0.45	0
BERT w/c	0.3	0.069	0.454	0.005
BERT w/c frozen	0.318	0.008	0.452	0.004
EmoRoBERTa	0.09	0	0.45	0
EmoRoBERTa frozen	0.09	0	0.45	0
EmoRoBERTa w/c	0.314	0.076	0.45	0
EmoRoBERTa w/c frozen	0.348	0.03	0.45	0

Conclusion

- Study Overview:
  - Total of six models evaluated, including four BERT models in frozen and unfrozen configurations, resulting in ten models for assessment.
- Performance Evaluation:
  - Majority classifier effective only for '0' trigger and neutral emotion categories, while random classifier exhibited average performance.
  - BERT-based models, particularly unfrozen ones, consistently outperformed frozen counterparts, leveraging training of embedding layers.
- Contextual Analysis:
  - Observation: Accuracy improved inversely proportional to the amount of context provided as input.
  - EmoRoBERTa models: Non-contextual model outperformed contextual counterpart, with frozen configuration yielding better results.
- Implications:
  - Conclusion: Context is less crucial than expected for the task. Treating sentences as more independent improves accuracy.
  - Potential limitation: Context may lead to overfitting, while simplified models facilitate faster convergence of loss function.
- Future Directions:
  - Explore advanced model architectures capable of handling larger contextual information efficiently.
  - Balance between context utilization and computational efficiency for improved performance.