Multimodal emotion recognition

Communicative Robots | Fall 2020
VU Amsterdam

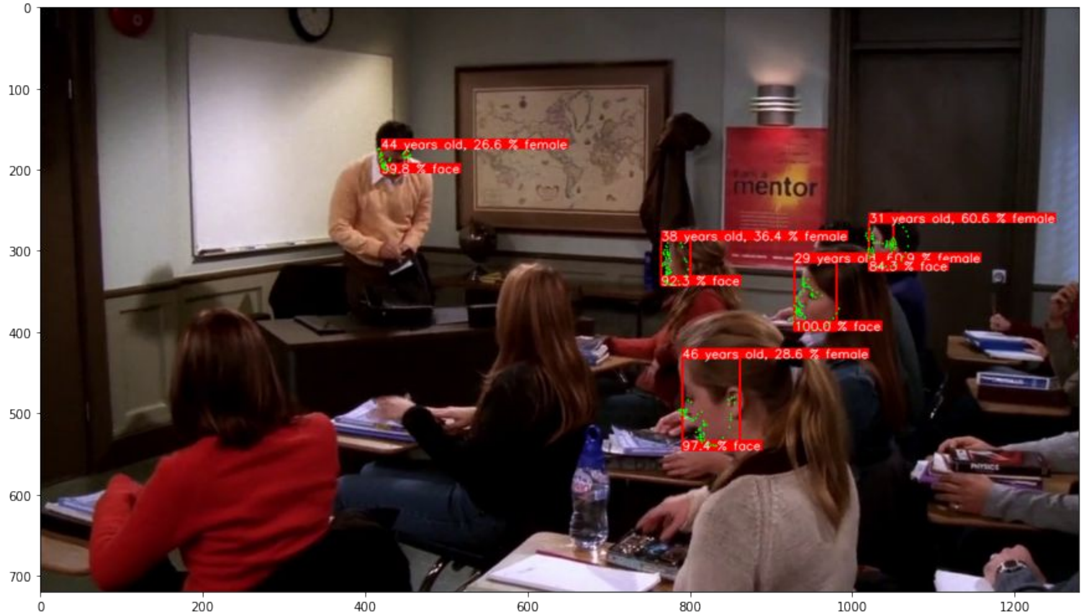
Contents

1. Train / dev / test results per modality:
    a. Vision (Tae) - 2 mins
    b. Text (Nihat, Zeynep) - 5 mins
    c. Audio (Vivian) - 5 mins
2. Modality fusion (Zeynep) - 3 mins
3. Critical analysis of complex emotions in the data (Wesley) - 5 mins
4. Discussion - 10 mins

1.a. Train / dev / test results on vision

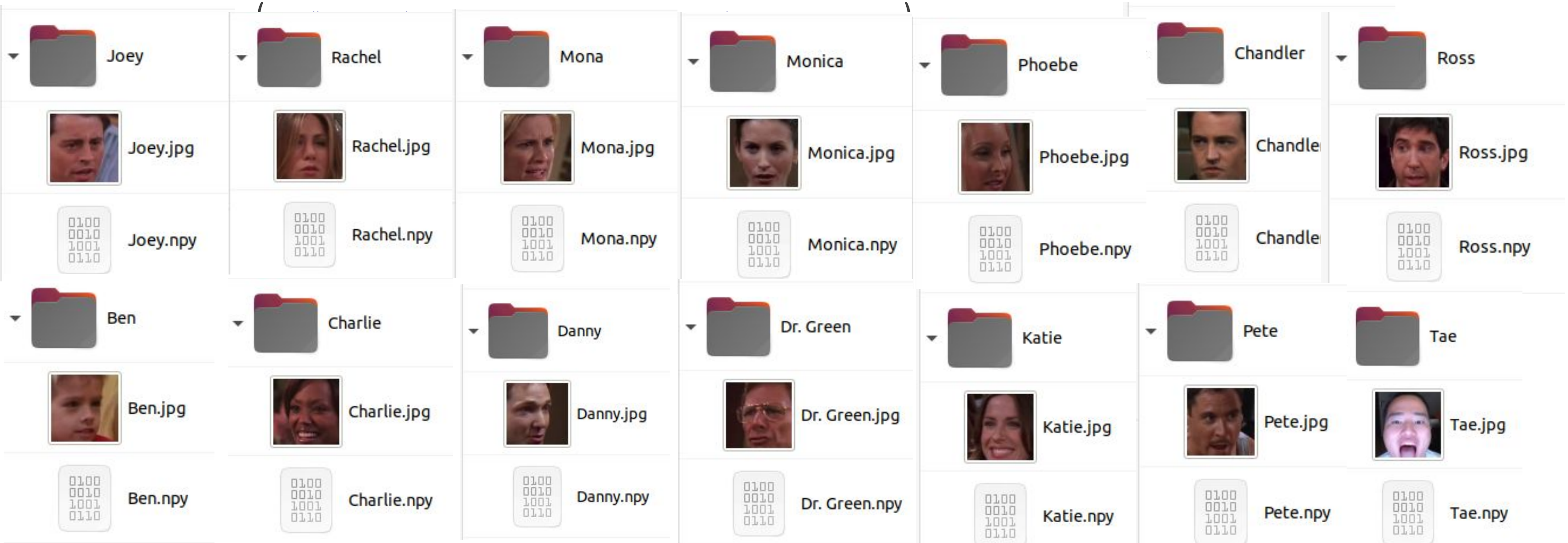
- There are two problems:
  - The videos are not perfectly aligned with text.
    - Data is always supposed to be dirty. I can live with that.
  - **We don’t know where the speaker is in the given frame.**
    - Is he/she even there?
    - What if there are multiple faces detected?

1.a. Train / dev / test results on vision



1.a. Train / dev / test results on vision

- Run face recognition on a smaller dataset



1.a. Train / dev / test results on vision

- Only extract the faces that match the speaker
- Extract landmarks
- Example: https://youtu.be/uX3g5NgvIj8 (disgust)
- Train with a 2 layered bidirectional LSTM
- The results are awful!
  - https://github.com/leolani/cltl-face-all/blob/master/examples/colab/4.ERC-MELD-compact-visual-colab.ipynb
  - slightly better than classifying everything as “neutral”
  - label: *test emotions {'surprise': 0.167, 'neutral': 0.509, 'anger': 0.056, 'sadness': 0.12, 'joy': 0.13, 'fear': 0.009, 'disgust': 0.009}*

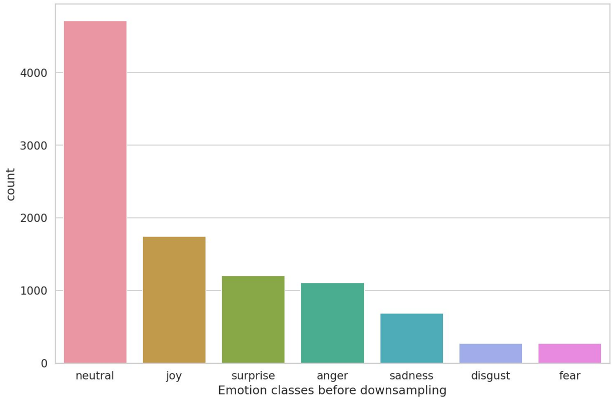
1.b. Text Classification on Friends Dataset

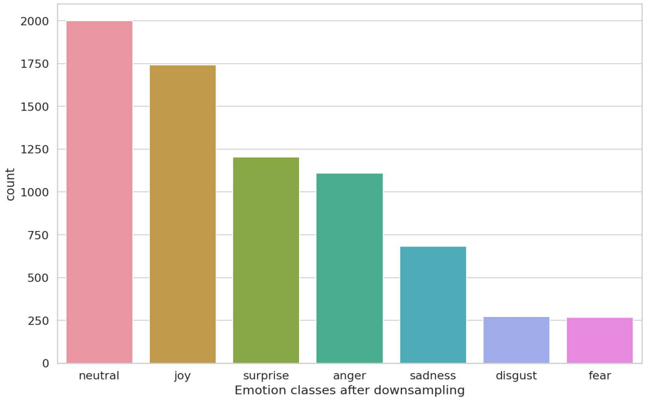
Column Specification

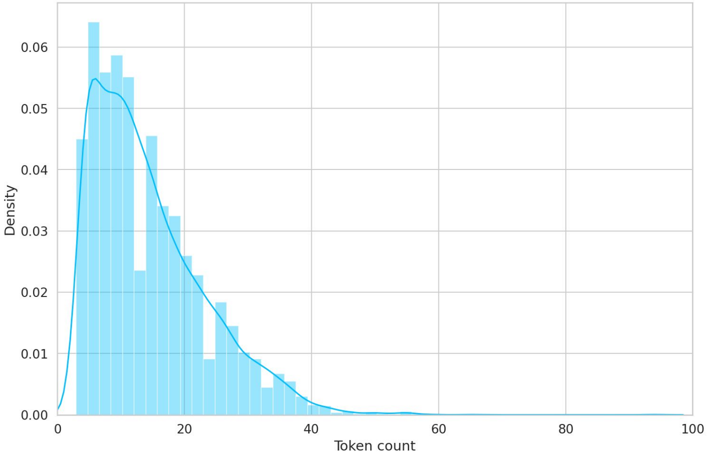
Column Name	Description
Sr No.	Serial numbers of the utterances mainly for referencing the utterances in case of different versions or multiple copies with different subsets
Utterance	Individual utterances from EmotionLines as a string.
Speaker	Name of the speaker associated with the utterance.
Emotion	The emotion (neutral, joy, sadness, anger, surprise, fear, disgust) expressed by the speaker in the utterance.
Sentiment	The sentiment (positive, neutral, negative) expressed by the speaker in the utterance.
Dialogue_ID	The index of the dialogue starting from 0.
Utterance_ID	The index of the particular utterance in the dialogue starting from 0.
Season	The season no. of Friends TV Show to which a particular utterance belongs.
Episode	The episode no. of Friends TV Show in a particular season to which the utterance belongs.
StartTime	The starting time of the utterance in the given episode in the format 'hh:mm:ss,ms'.
EndTime	The ending time of the utterance in the given episode in the format 'hh:mm:ss,ms'.

*https://github.com/declare-lab/MELD

1.b. Downsampling & Sequence Length







- TRAIN Dataset: 9989 -> 7279,
- VALIDATION Dataset: 1109,
- TEST Dataset: 2610
  - neutral 1256
  - joy 402
  - anger 345
  - surprise 281
  - sadness 208
  - disgust 68
  - fear 50

1.b. Text Classification on Friends Dataset

Model Name	Base Model	Pooling	Training Data	STSb Performance (Higher = Better)
roberta-large-nli-stsb-mean-tokens	roberta-large	Mean Pooling	NLI+STSb	86,39
roberta-base-nli-stsb-mean-tokens	**roberta-base**	Mean Pooling	NLI+STSb	**85,44**
bert-large-nli-stsb-mean-tokens	bert-large-uncased	Mean Pooling	NLI+STSb	85,29
distilbert-base-nli-stsb-mean-tokens	distilbert-base-uncased	Mean Pooling	NLI+STSb	85,16
bert-base-nli-stsb-mean-tokens	bert-base-uncased	Mean Pooling	NLI+STSb	85,14

- 12-layer, 768-hidden, 12-heads, 125M parameters
- RoBERTa using the BERT-base architecture

*https://github.com/UKPLab/sentence-transformers
*https://github.com/pytorch/fairseq/tree/master/examples/roberta

1.b. Custom Roberta Model

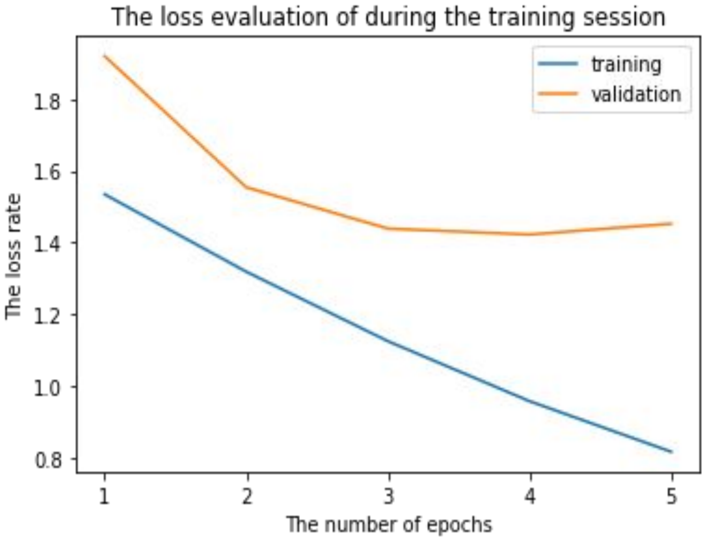
- Pretrained Roberta Base -> Linear -> ReLU -> 30% Dropout -> Linear -> Output (7 classes)
- Cross Entropy Loss
  - Combination of LogSoftmax and Negative Log Likelihood

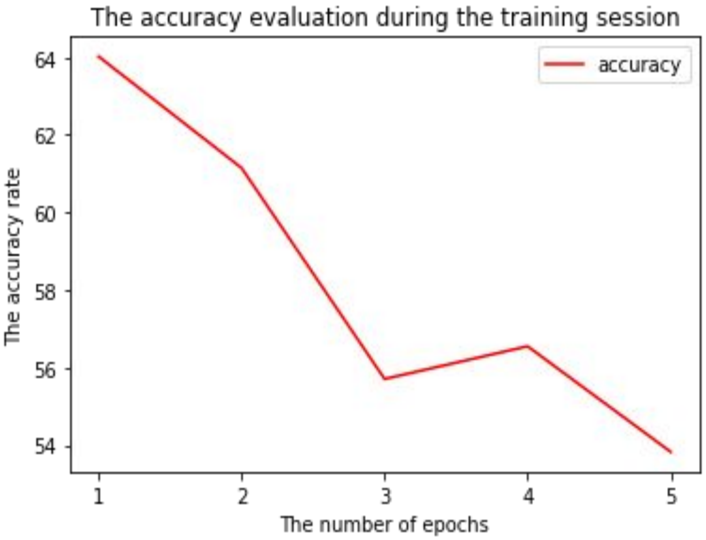
*https://arxiv.org/abs/1207.0580

1.b. Text Classification on Friends Dataset

Model Name	Accuracy	**weighted F1-score**
Fully training in Roberta	64.02% (max 64.21%)	61%
Roberta Transformer + SVM	57.1%	57.2%
Roberta Transformer + XGBoost Classifier	60%	58%
Roberta Transformer + Logistic Regression	58%	58%

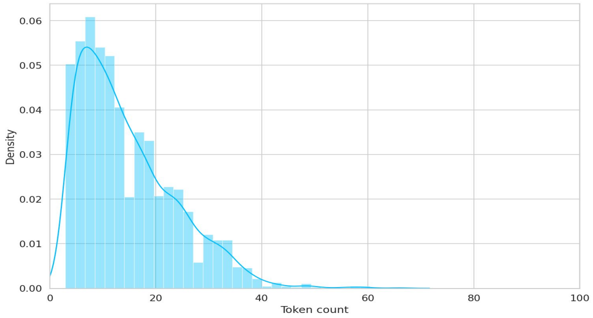
1.b. Text Classification on Friends Dataset





1.b. Text Classification on Friends Dataset

Take aways;

- Random downsampling does not help on accuracy in a sequential dataset
  - Got 1.4% higher rate from ordered then downsampled version
- Sequence length choice matters
  - 50 as an embedding size dropped about 3% in accuracy
  - 

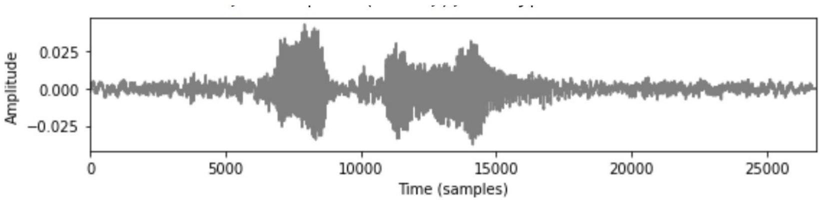
longest: 66 tokens in test dataset

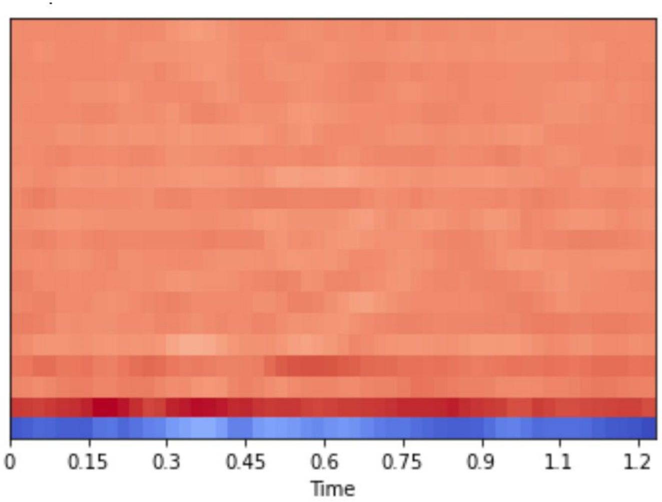
- Pretrained models don’t need many rounds of training

1c. Audio - Features

Librosa package to extract audio features from given mp3 files.

- Mel Scale (pitch)
- MFCC
- Chroma (pitch)
- Zero crossing rate
- Spectral rolloff (frequency)
- Spectral centroid (“brightness”)
- Spectral bandwidth
- RMSE (loudness)





1c. Audio - Final Data

- 185 features (mean for each feature in each audio file)
- Normalised data with MinMaxScaler

- 2 datasets prepared:
  - Original dataset
  - Downsampled dataset where the largest class (neutral) is of the same size as second-largest class (joy)
  - (Fully balanced did not provide enough examples per class (256 only))

1c. Audio - Model set-up

- Tested two neural networks
  - PyTorch simple linear model
  - Keras fully-connected deep neural network (2 hidden layers)
- Experimented with different settings (epochs, batch size, dropout rate, etc.)

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 185)               34410
_________________________________________________________________
dropout_3 (Dropout)          (None, 185)               0
_________________________________________________________________
dense_5 (Dense)              (None, 128)               23808
_________________________________________________________________
dropout_4 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_6 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_5 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_7 (Dense)              (None, 7)                 455
=================================================================
Total params: 66,929
Trainable params: 66,929
Non-trainable params: 0
_________________________________________________________________
```

1c. Audio - Results

The results overall are not too great...

- Original data
  - Linear model: 20-25% accuracy
  - Fully-connected: 27% accuracy
- Down-sampled data:
  - Linear model: 22-28%
  - Fully-connected: **30%**

Downsampled dataset

	precision	recall	f1-score	support
0	0.20	0.57	0.30	345
1	0.00	0.00	0.00	68
2	0.00	0.00	0.00	50
3	0.17	0.26	0.20	402
4	0.58	0.33	0.42	1256
5	0.00	0.00	0.00	208
6	0.19	0.18	0.18	281
accuracy			0.30	2610
macro avg	0.16	0.19	0.16	2610
weighted avg	0.35	0.30	0.29	2610

Original dataset

*************** test ***************

	precision	recall	f1-score	support
0	0.23	0.36	0.28	345
1	0.00	0.00	0.00	68
2	0.00	0.00	0.00	50
3	0.16	0.55	0.25	402
4	0.55	0.27	0.37	1256
5	0.00	0.00	0.00	208
6	0.27	0.07	0.11	281
accuracy			0.27	2610
macro avg	0.17	0.18	0.14	2610
weighted avg	0.35	0.27	0.26	2610

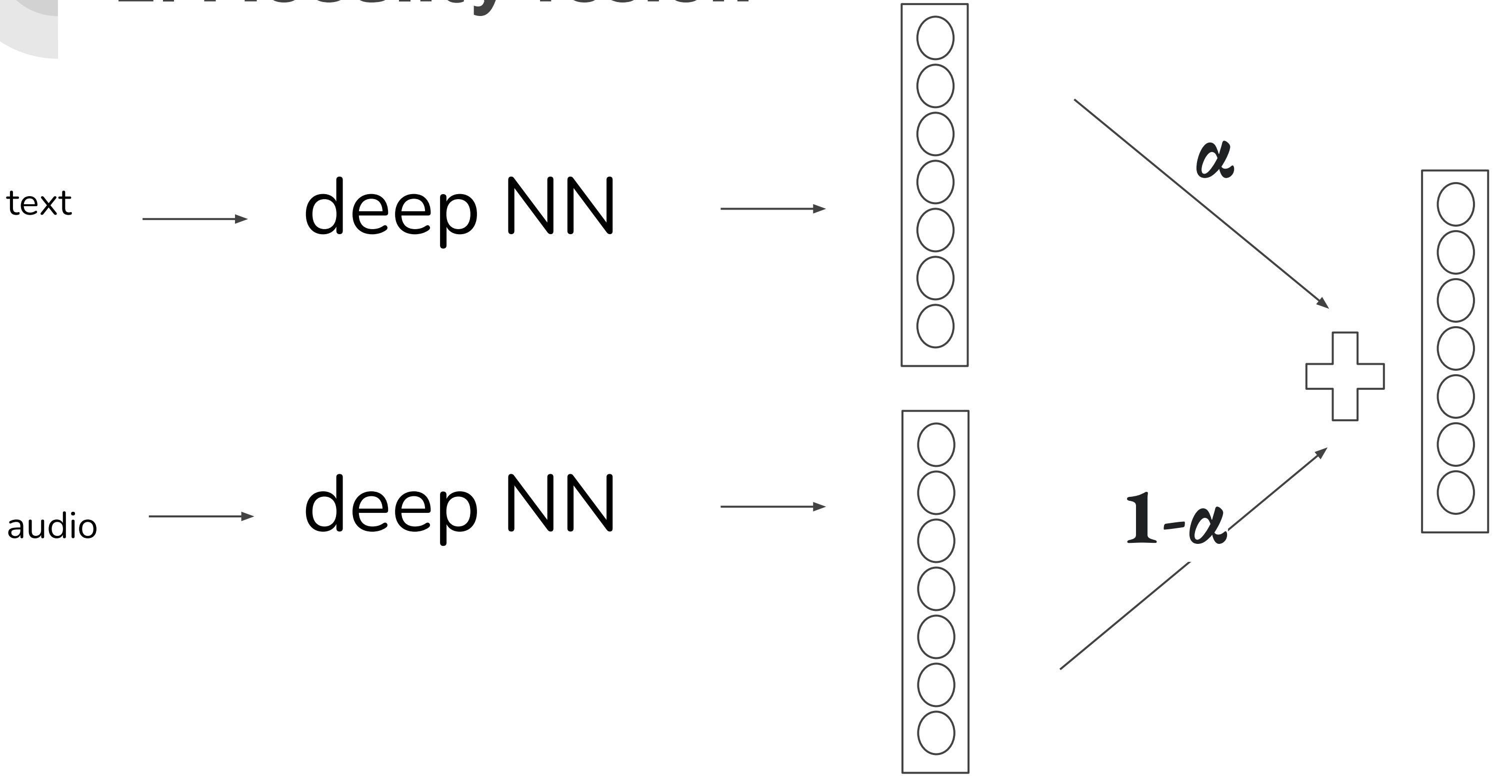
1c. Audio - Explanations

- Possible explanations for why the model does not perform well:
  - Model overfit (validation loss increasing) → solved by Dropout
  - Something is wrong with the NN → have tested different setups and overall accuracy did not improve
  - Emotion extraction out of audio alone is just really difficult
  - MELD authors use openSMILE and report a weighted-average accuracy between 39-42 for an LSTM and RNN.

2. Modality fusion

- Obviously we can’t use vision for the mentioned problems.
- We’ll merge audio and text.
- Several ways of fusion:
    - Early fusion, where the raw audio and text data are merged in the early layers of neural networks
        - This is not easy to optimize.
        - Needs more time.
    - Late fusion, where the audio and text modalities meet later part of the layers.
        - This is easier
        - It’s even easier when the modalities are trained separately.

2. Modality fusion



2. Modality fusion

alpha is found to be the best when it's 0.7. This means that we put 70% emphasis on text and the other 30% on audio. One can also consider this value to be "attention"

weighted f1 score (see: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

	audio only	text only	audio+text	COSMIC (SOTA)
train	0.373	0.657	0.656	
dev	0.252	0.578	0.587	
test	0.267	0.617	**0.619**	0.652

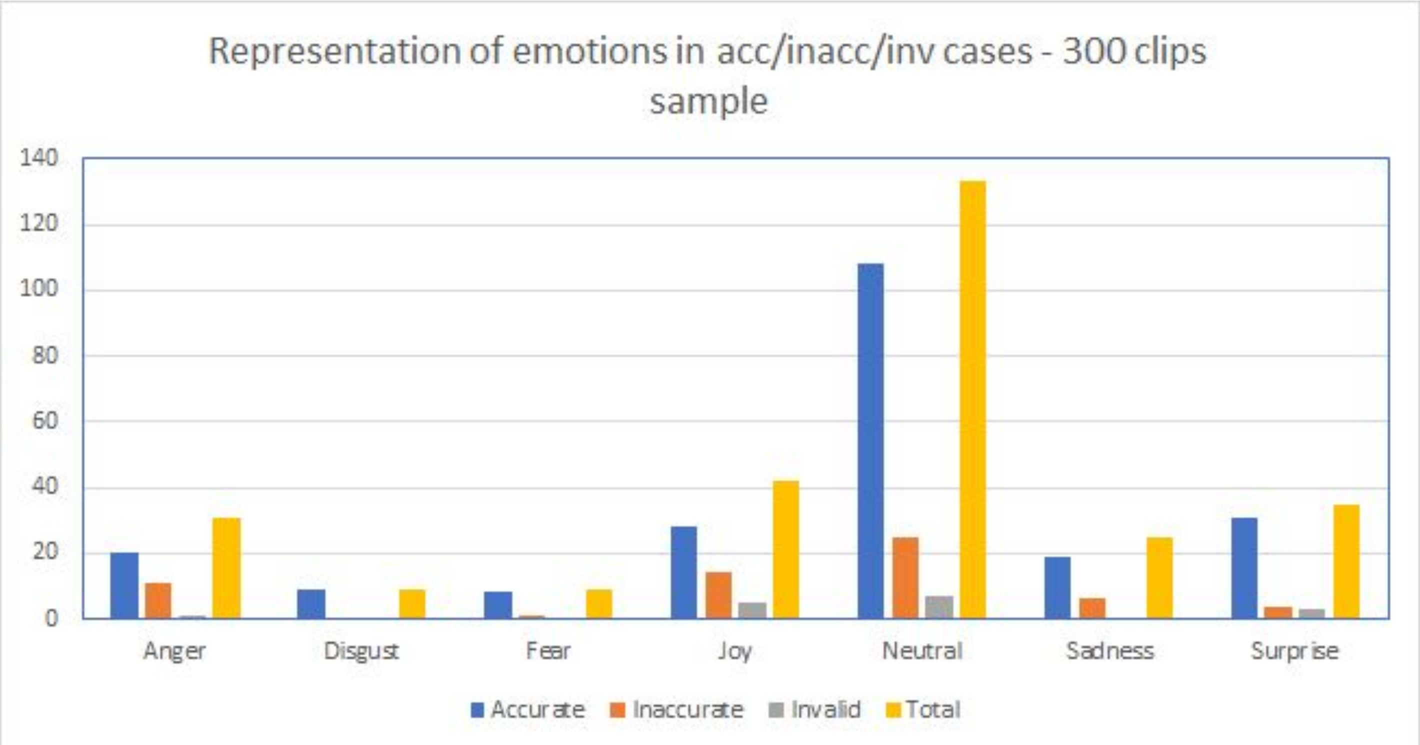
Actually we did pretty good. The fusion worked!

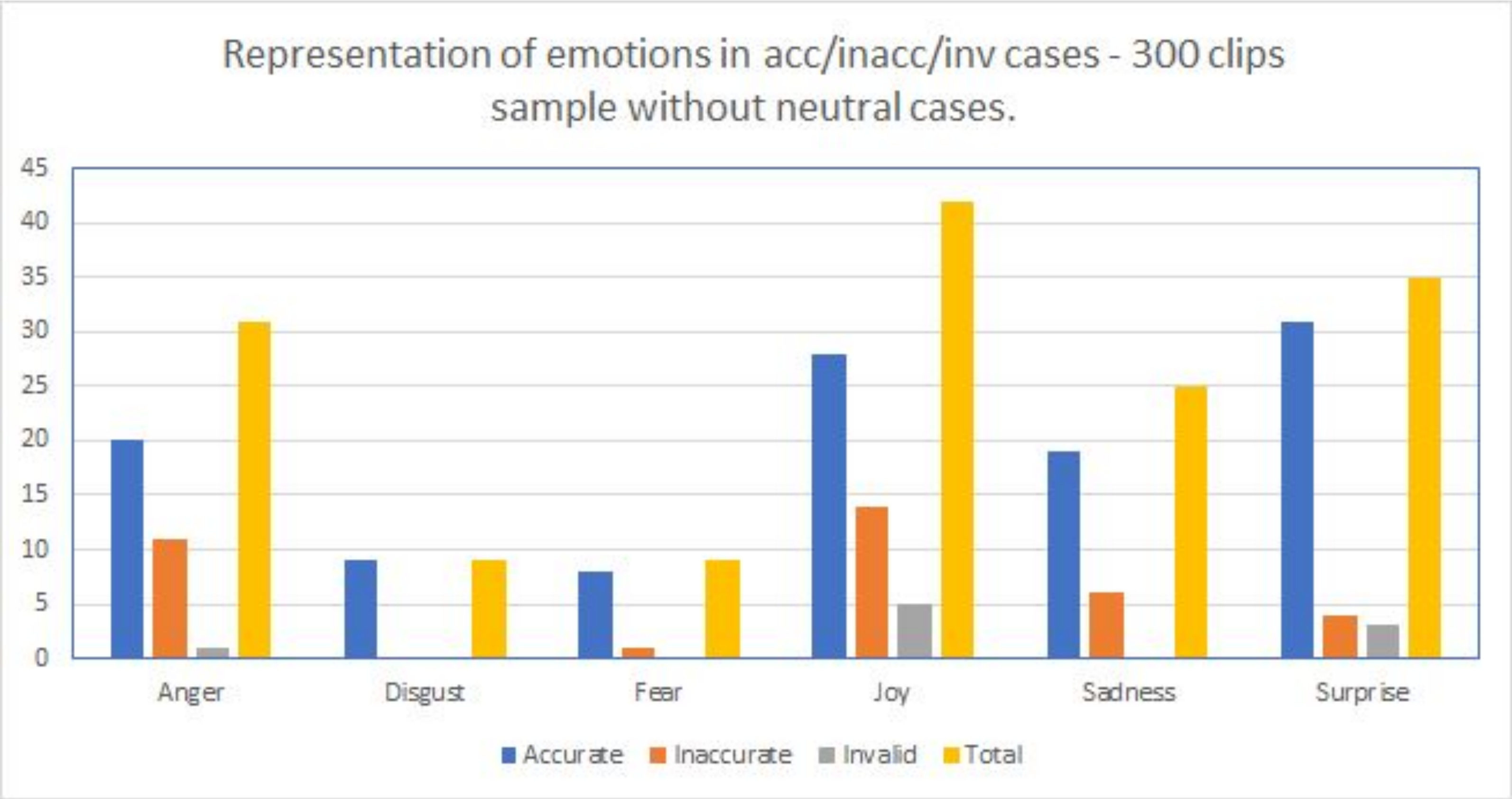
3. Critical analysis of complex emotions in the data

- 300 clips from Friends
- Analyzed by checking the facial landmarks for each relevant frame
- Proceeding with checking whether the basic emotions follow Ekman’s theorem (facial features displayed per emotion) in the clip
- If yes, no additional comment
- If no, elucidations!

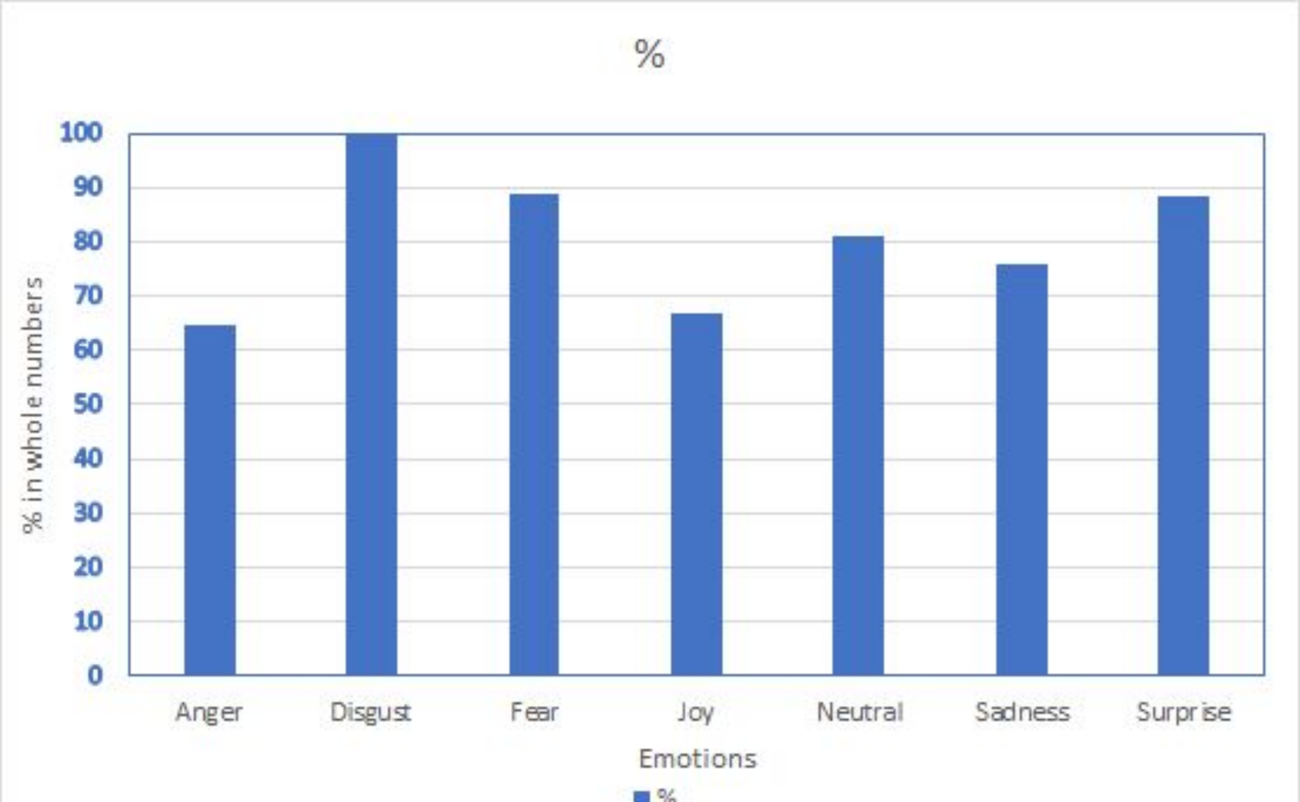
Anger cli	# of frame	Person with emotio	Landmark features	Aligns with theorem
3	82	Chandler	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes
7	29	Rachel	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, neutral expression is shown
8	26	Rachel	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, neutral expression is shown*
13	115	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, neutral expression/slight sadness (see 12)
72	46	Phoebe	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, slightly audible in voice, though.
84	89	Rachel	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
90	46	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes. The clip overlaps with 89, showing that it did detect the anger.
112	114	Phoebe	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, but it is yet again slightly audible in voice.
113	70	Monica	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes. Emotion is more audible than visible, though.
122	41	Richard	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
125	62	Joey	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
126	82	Richard	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
127	153	Joey	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes
166	57	Katie	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes, however, it's more audible than visible due to angles.
185	129	Dr. Green	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
187	43	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
189	49	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes, his facial landmarks show it.
190	53	Dr. Green	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes, but the angle shows one side of his face so the data is unreliable.
193	7	Dr. Green	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
195	91	Dr. Green	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes, surprise in the first few frames however.
199	50	Dr. Green	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
200	24	Chandler	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, more signs of disgust and a hint of sadness rather than anger.
220	324	Phoebe	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, more signs of a neutral expression.
238	13	Chandler	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, covers his eyes and there's not a good angle to it.
241	65	Monica	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, there are no clear signs of anger on Monica's face.
244	53	Monica	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
275	34	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
277	69	Ross	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
288	44	Rachel	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, Rachel seems to be more confused by something than anything else.
289	13	Rachel	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
290	30	Phoebe	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	Yes.
299	28	Phoebe	Eyebrows pulled down, squinted eyes, lips rolled in and they may be tightene	No, Phoebe doesn't show a single sign of anger at all, more neutrality than anything else.

3. Critical analysis of complex emotions in the data




Representation of emotions in acc/inacc/inv cases - 300 clips sample without neutral cases.
45
40
35
30
25
20
15
10
5
0
Anger
Disgust
Fear
Joy
Sadness
Surprise
Accurate
Inaccurate
Invalid
Total

3. Critical analysis of complex emotions in the data



3. Critical analysis of complex emotions in the data.

And a brief explanation as to why the data is the way it is.

- Data took text, voice, and facial features in consideration
- Micro-expressions/emotions
- Ekman’s basic emotions model
- They’re actors!

4. Discussion

- Our codes can be found at: https://github.com/leolani/cltl-face-all/tree/master/examples/colab