Multimodal Speech Emotion Recognition Using Audio and Text

Group J: Jan Arvin Lapuz, Alyssa Lim, Le Van Nguyen, Ramil Zabala

COMP8240 Applications of Data Science, April 2014

Paper Abstract

- Speech Emotion Prediction from combined audio and text
- Model - audio and text
- Environment - Python 2.7, Colab Tensorflow 1.4
- Github link to the orig data

Replication Results

Model	Replication Accuracy	Research Accuracy	
Text Only	0.00%	63.5%	62.8%
Audio Only	51.1%	54.6%	62.8%
Multimodal	0.00%	71.0%	71.8%
Multimodal-Attention	49.9%	69.0%	62.8%

New Data - Youtube Audio and Transcript

Audio

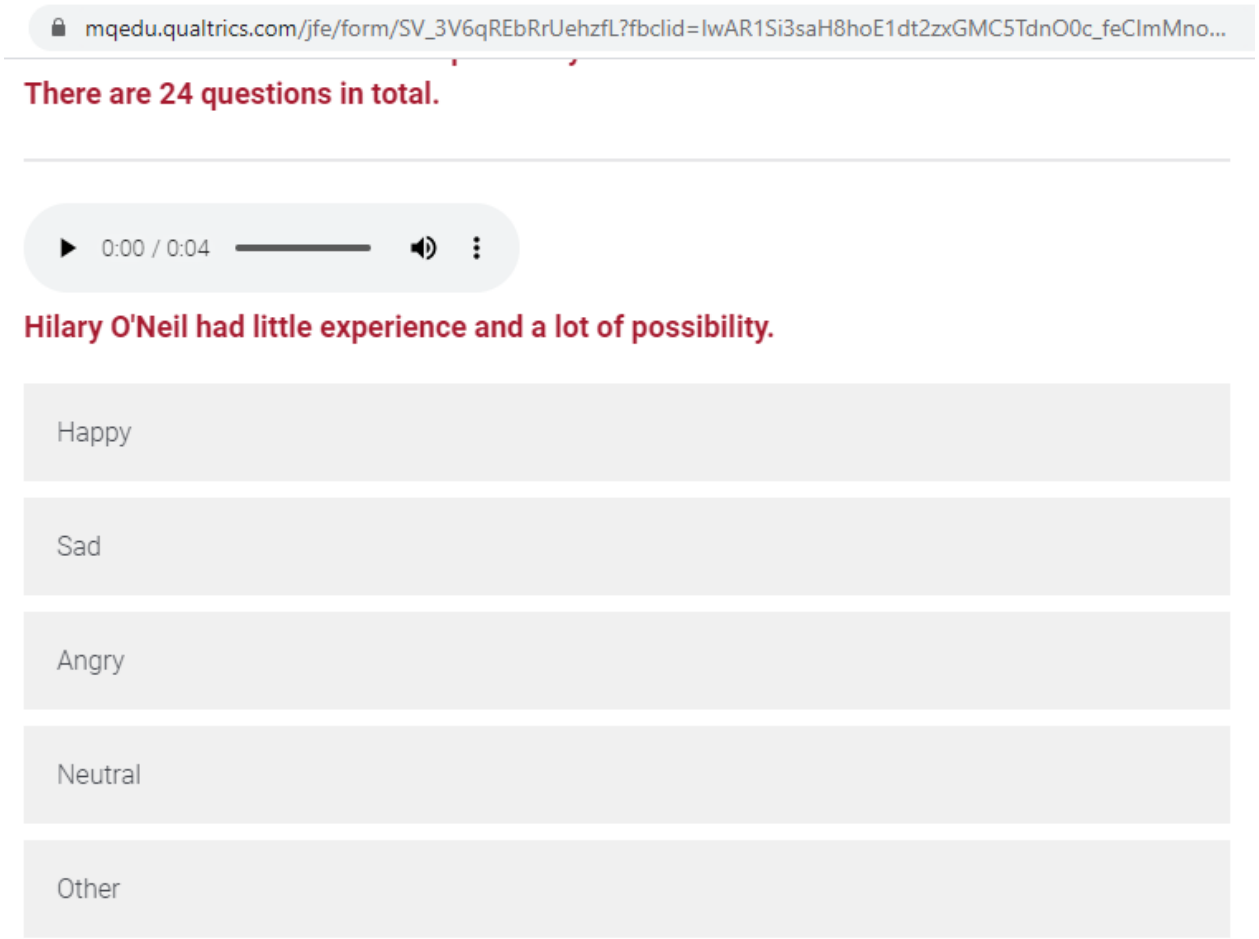
- Youtube_dl Python package - extract the audio
- Sox package - mp3 to wav conversion

Transcript

- savesubs.com to save the captions
- Autogenerated captions

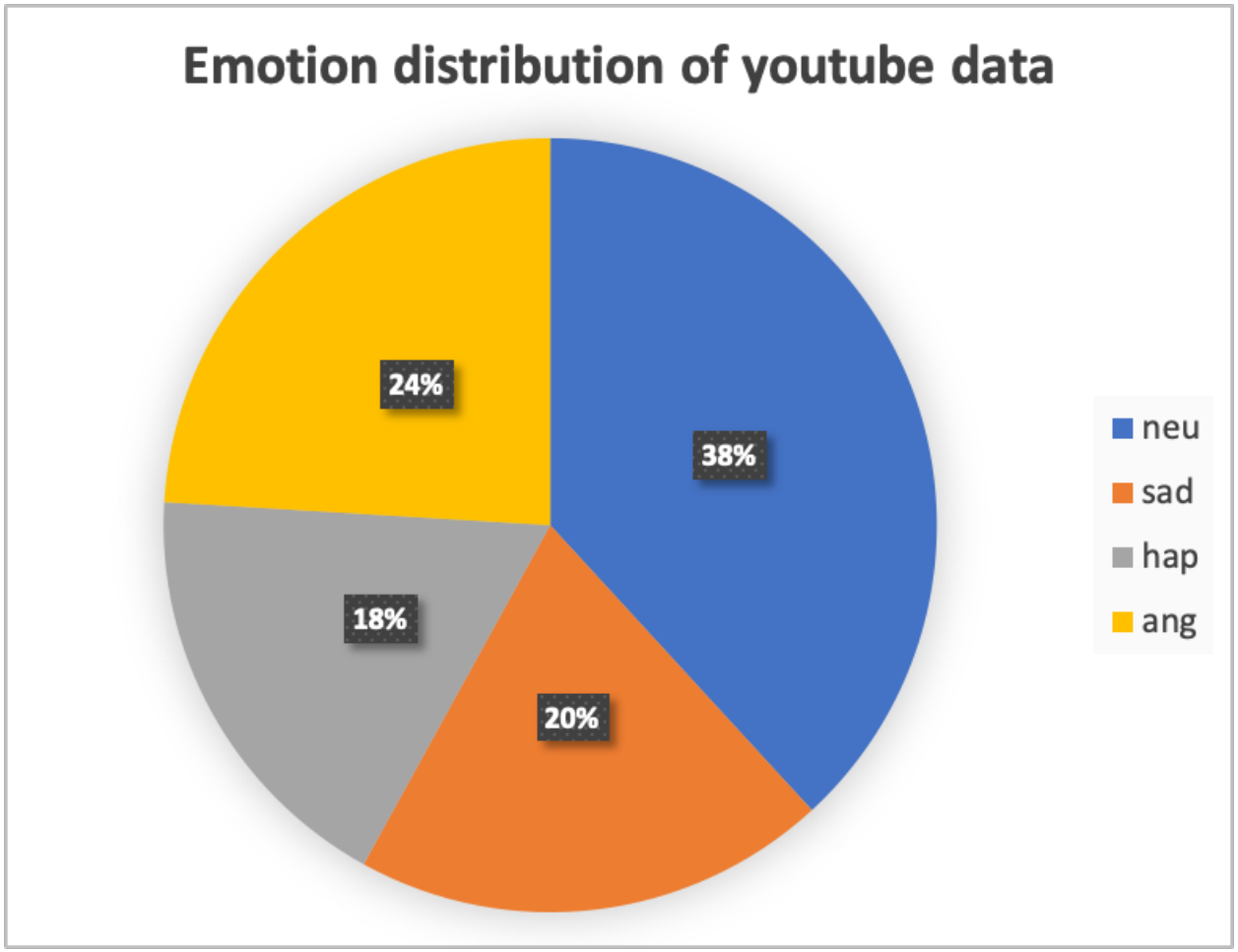
New Data - Youtube Emotion

Qualtrics Emotion Annotation Survey



New Data - Emotion Distribution

Emotion Distribution 284 Transcripts Collected



New Data - TESS data

- Attach the structure of the test data – Tess data structure mismatch
- Get description of the dataset
- Google API to get the transcript
- Mention the research article
- Tess data structure mismatch

Movie snippet

alt content

New Data Preprocessing - MFCC

MFCC (Audio Features)

- Extract features through Open Smile
- Major MFCC Features
  - Volume
  - Energy
  - Pitch
  - Zero Crossing Rate
  - Spectral Centroid

	keyfld	v1	v2	v3	v4	v5	v6	v7	v8	v9	...	v31	v32	v33	v34	
0	'unknown'	0.00	-6.109069	-14.02484	6.732327	-25.465800	6.154059	-17.885570	7.608350	-21.891620	...	0.286751	0.432377	0.723979	0.882882	0
1	'unknown'	0.01	-3.712081	-19.80455	10.101170	-21.713700	-2.548810	-25.203900	0.034481	-23.605340	...	0.367582	0.859231	1.380296	1.035713	-0
2	'unknown'	0.02	-2.756485	-16.49574	13.446880	-21.667230	4.936851	-21.908060	9.636816	-7.501993	...	-0.053051	0.702943	1.619453	0.298885	-1
3	'unknown'	0.03	-1.430273	-14.00117	8.582673	-17.582100	4.218672	-21.558180	18.152650	-0.833221	...	-0.759775	0.402543	0.893972	-1.149457	-1
4	'unknown'	0.04	-0.820436	-12.82985	6.748813	-16.742400	6.498775	-10.289510	12.204360	-25.371730	...	-0.838942	-0.279553	-0.817221	-1.662912	0
5	'unknown'	0.05	0.471009	-14.57329	8.681200	-11.869240	7.933800	-12.737730	15.178280	-17.512030	...	-0.316393	-0.413891	-1.453950	-1.281607	0
6	'unknown'	0.06	-4.287149	-22.79697	-2.673139	-21.437270	8.180445	-5.493269	14.403130	-9.281401	...	0.462294	0.171155	-0.987680	0.271508	0

New Data Preprocessing - Prosody

Prosody (Emotion Features)

- Extract features through Open Smile
- Major Prosody Features
  - Intonation
  - Stress
  - Rhythm
  - Speech Rate
  - Reduced Forms and Linking
  - Pauses
  - Voice Quality

	keyfld	v1	v2	v3	v4	v5	v6	v7	v8	v9	...	v31	v32	v33	v34	
0	'unknown'	0.00	-6.109069	-14.02484	6.732327	-25.465800	6.154059	-17.885570	7.608350	-21.891620	...	0.286751	0.432377	0.723979	0.882882	0
1	'unknown'	0.01	-3.712081	-19.80455	10.101170	-21.713700	-2.548810	-25.203900	0.034481	-23.605340	...	0.367582	0.859231	1.380296	1.035713	-0
2	'unknown'	0.02	-2.756485	-16.49574	13.446880	-21.667230	4.936851	-21.908060	9.636816	-7.501993	...	-0.053051	0.702943	1.619453	0.298885	-1
3	'unknown'	0.03	-1.430273	-14.00117	8.582673	-17.582100	4.218672	-21.558180	18.152650	-0.833221	...	-0.759775	0.402543	0.893972	-1.149457	-1
4	'unknown'	0.04	-0.820436	-12.82985	6.748813	-16.742400	6.498775	-10.289510	12.204360	-25.371730	...	-0.838942	-0.279553	-0.817221	-1.662912	0
5	'unknown'	0.05	0.471009	-14.57329	8.681200	-11.869240	7.933800	-12.737730	15.178280	-17.512030	...	-0.316393	-0.413891	-1.453950	-1.281607	0
6	'unknown'	0.06	-4.287149	-22.79697	-2.673139	-21.437270	8.180445	-5.493269	14.403130	-9.281401	...	0.462294	0.171155	-0.987680	0.271508	0

New Data Model Testing

- Absence of pre-trained model
- Retraining on original IEMOCAP data - results close to the original
- Evaluation and Training on New Data
- Small modifications to the configuration

Summary and Reflections

- Replication results
- Effect of the Tess dataset variability
- Weights Parameters in the Open Smile
- Effect of music background
- Controlled environment in the IEMOCAP data