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, Nov 2020

Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

Paper Abstract

Paper:

- Multimodal Speech Emotion Recognition Using Audio and Text by Seunghyun Yoon, Seokhyun Byun, Kyomin Jung
- 2018 IEEE Spoken Language Technology Workshop (SLT)
- Rank 14 in Computational Linguistics Category on Google Scholar
- Source code is available on author's Github repository without the pretrained model

Four models: Text only, Audio only, Multimodal, Multimodal-Attention

Processed input data:

- Speech Data (.npy file): Mel Frequency Crystal Coefficients (MFCC), Prosodic Features
- Text Data (.npy file): Word Tokens
- Emotion Categories: Angry, Happy, Sad, Neutral

Set up for Google Colab:

- tensorflow==1.4; python==2.7
- scikit-learn==0.20.0; nltk==3.3

Original Replication Results

Replication of the Original Data (IEMOCAP data)

Model	Published Acc	Replication Acc
Text Only	63.5%	62.8%
Audio Only	54.6%	55.7%
Multimodal	71.8%	71.0%
Multimodal-Attention	69.0%	48.5%

Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

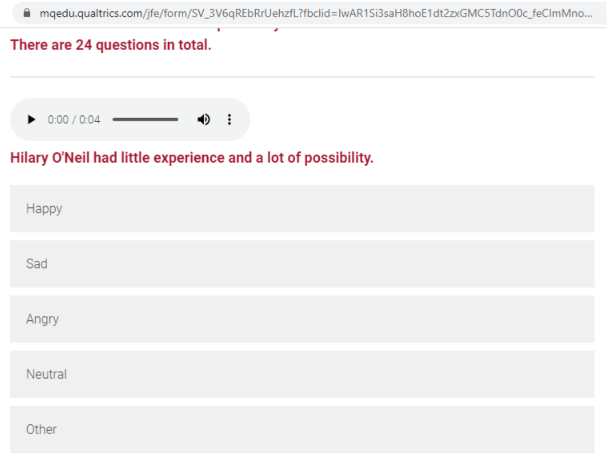
Data Collection: YouTube

Approach 1: YouTube

- Data Sources:
    - Dying Young Trailer
      `https://www.youtube.com/watch?v=A8p0w_Ec1NY/`
    - Trailers of Top 5 romance movies of all time
      `https://www.youtube.com/watch?v=y0F_JE-dSxg/`
    - UNTUCKED: Rupaul's Drag Race S09 E04
      `https://www.youtube.com/watch?v=6-Eg_TaGfTI/`
- Audio Extraction:
    - Used youtube_dl package in Python
    - Converted from mp3 to wav using the sox package
- Transcription:
    - Auto-generated captions from YouTube extracted through savesubs.com

Data Annotation: YouTube

Qualtrics Emotion Annotation Survey



mqedu.qualtrics.com/jfe/form/SV_3V6qREbRrUehzfL?fbclid=IwAR1Si3saH8hoE1dt2zxGMC5TdnO0c_feClmMno...

There are 24 questions in total.

Hilary O'Neil had little experience and a lot of possibility.

Happy

Sad

Angry

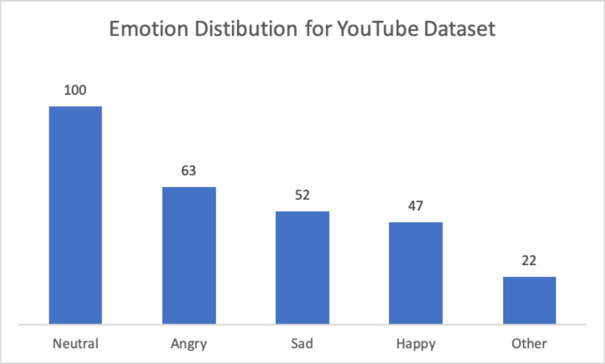
Neutral

Other

Data Annotation Distribution

YouTube Emotion Distribution

284 Transcripts Collected in Qualtrics



Research Data: TESS

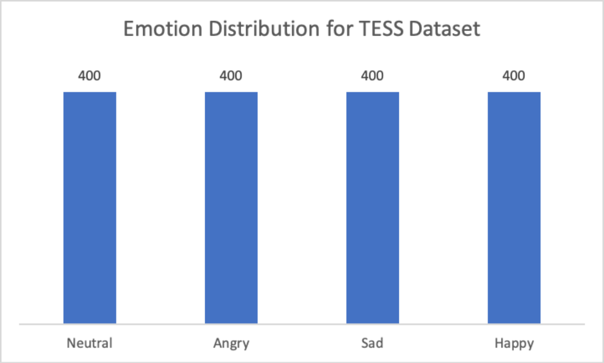
Approach 2: Toronto Emotional Speech Test

- Audio Description:
  - Collected by the University of Toronto, Psychology Department in 2010. There were 200 target words were spoken by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions
  - Filtered the emotions under category: anger, happiness, sadness, neutral
- Transcription:
  - Google Speech API was used to extract the transcript

TESS Emotion Distribution

TESS Emotion Distribution

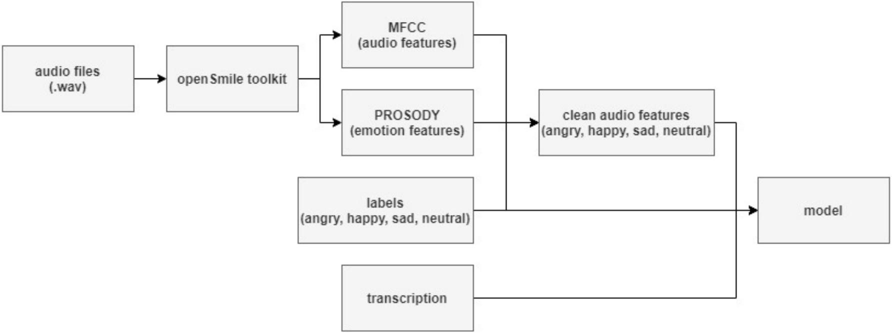
8 sets (2 speakers, 4 emotions) with 200 Transcripts each



Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

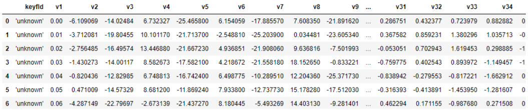
New Data Preprocessing: Diagram



New Data Preprocessing: Feature Details

MFCC (39 Features)

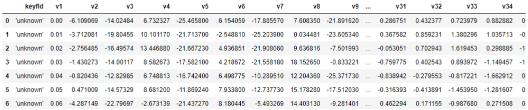
- Volume
- Energy
- Pitch
- Zero Crossing Rate
- Spectral Centroid



	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

Prosody (35 Features)

- Intonation
- Stress
- Rhythm
- Speech Rate
- Pauses
- Voice Quality



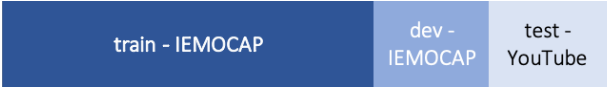
	keyfld	v1	v2	v3	v4	v5	v6	v7	v8	v9	...	v31	v32	v33	v34	
0	'unknown'	0.00	-6.109069	-14.02484	6.732327	-25.465800	6.154059	-17.885570	7.608350	-21.891620	...	0.286751	0.432377	0.723979	0.882882	0
1	'unknown'	0.01	-3.712081	-19.80455	10.101170	-21.713700	-2.548810	-25.203900	0.034481	-23.605340	...	0.367582	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

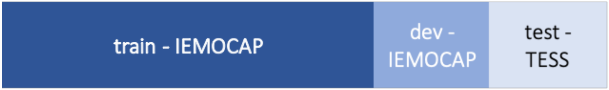
Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

Testing the Model on New Data

- Absence of pretrained model
- Retraining on original IEMOCAP data - results close to the original
- Training on IEMOCAP data and evaluating the model on the new data (YouTube and TESS)





- Small modifications to the configuration

Presentation Outline

1. Project Overview and Original Replication Results
2. New Data Collection and Annotation
3. New Data Preprocessing
4. Testing the Model on New Data
5. Replication Results and Reflection

Replication Results

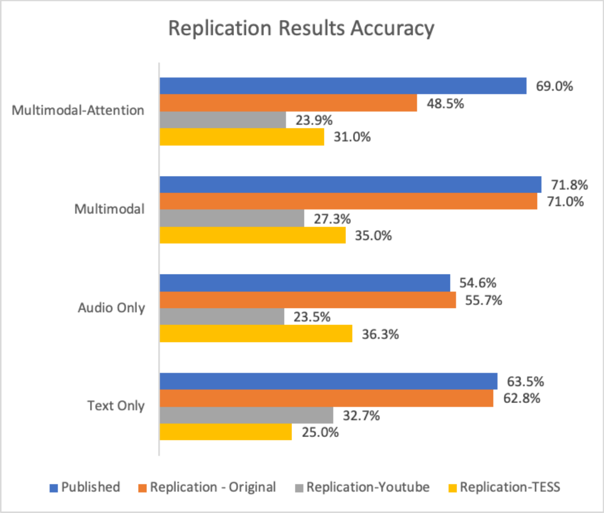
Replication Results on the Original Data

Model	Published Acc	Replication Acc
Text Only	63.5%	62.8%
Audio Only	54.6%	55.7%
Multimodal	71.8%	71.0%
Multimodal-Attention	69.0%	48.5%

Replication Results on the New Data

Model	Repl Acc YT	Repl Acc TESS
Text Only	32.7%	25.0%
Audio Only	23.5%	30.6%
Multimodal	27.3%	34.9%
Multimodal-Attention	23.9%	30.8%

Replication Results on Original and New Data



Summary and Reflections

- Effect of the TESS dataset variability, i.e., same transcripts spoken in different ways.
- Weight Parameters in openSMILE processing
- Effect of background music
- IEMOCAP data is collected in a controlled environment