Multimodal Speech Emotion Recognition Using Audio and Text [Research Replication]

Jan Arvin Lapuz, Alyssa Lim, Le Van Nguyen, Ramil Zabala

Abstract

A few studies have been carried out on developing models that classify emotions out of speech audio but never has dialogue transcript been incorporated to improve accuracy. In 2018, Seunghyun Yoon, et al. conducted a research that introduced a deep dual recurrent encoder model that combines audio features, such as MFCC and prosody, text data to predict the one of the four emotions angry, happy, sad and neutral. Their model achieved a WAP value of up to 71.8%. This paper attempted to replicate their results using the original IEMOCAP data and extend it to new datasets. The replication using the original data resulted in an accuracy of within 0.1% of the published result except for the MDREA model which had 21% reduction in accuracy. To extend the replicaton, new methodologies have been applied on YouTube videos to suit the model and reached a maximum WAP accuracy of 27%. Existing audio data such as the TESS dataset was also used with a slightly better maximum WAP accuracy of 31%.

1 Research Description

Since this project's purpose is to conduct an extended replication of a reproducible quality research, one paper that adapts these criteria is Multimodal Speech Emotion Recognition Using Audio and Text[1].

Many studies have been carried out that aim to create a model classifying emotion from speech. However, those models rely solely on audio features of speech data and does not deliver desired result. In the mentioned paper, Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung introduce a deep dual recurrent encoder model that simultaneously utilises both audio and transcript of speech data. In detail, the raw conversation dialogues and their transcripts are encoded by a dual recurrent neural networks (RNNs) before being combined to predict their emotion category - happy, sad, angry and neutral. By applying this architecture on the IEMOCAP dataset, the model is able to achieve up to 71.8% of accuracy.

For comparison, the paper considered four models in total:

- **Audio Recurrent Encoder(ARE)** utilises audio features only
- **Text Recurrent Encoder(TRE)** utilises text features only
- **Multimodal Dual Recurrent Encoder(MDRE)** utilises audio and text features simultaneously.
- **Multimodal Dual Recurrent Encoder with Attention(MDREA)** utilises audio and text features simultaneously with attention mechanism

The overall performance of the models are measured by the weighted average precision (WAP) value over the 5-fold cross validation data folders[1]. With this evaluation framework, the performance of MDRE is reported to have highest WAP in the paper (at around 71.8%).

[1] WAP calculation:
`https://github.com/david-yoon/multimodal-speech-emotion/blob/master/model/evaluation_text.py`

1.1 Justification

The quality of the paper is demonstrated by the fact that it is published as a part of the 2018 IEEE Spoken Language Technology Workshop (SLT), which is ranked 14th in Computational Linguistics Category on Google Scholar [**Google Scholar Computation linguistics**]. Furthermore, the paper was cited 58 times at the time of writing of this report[2]. Although a pre-trained model is not provided, the paper's source code is available through one of the authors' Github repository[3].

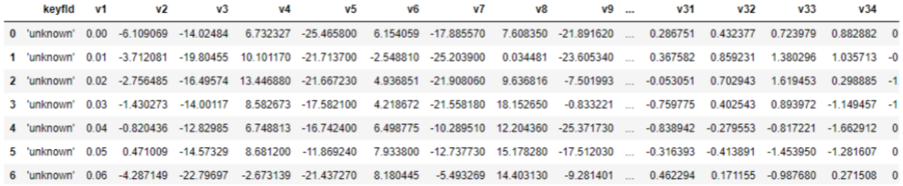
2 Original Dataset Description

The original dataset discussed in the paper is from Interactive Emotional Dyadic Motion Capture (IEMOCAP) database[4]. It consists of roughly 12 hours of audiovisual data of acted conversations from 12 actors in which only speech (.wav format) and transcript (.txt format) data are selected.

Before inputting into the architecture, some modifications had been made. The audio signal is extracted into Mel-frequency Cepstral Coefficients (MFCCs) and prosodic features using the openSMILE toolkit while the transcript is tokenised and indexed by Natural Language Toolkit(NLTK). The processed data are all in numpy format.

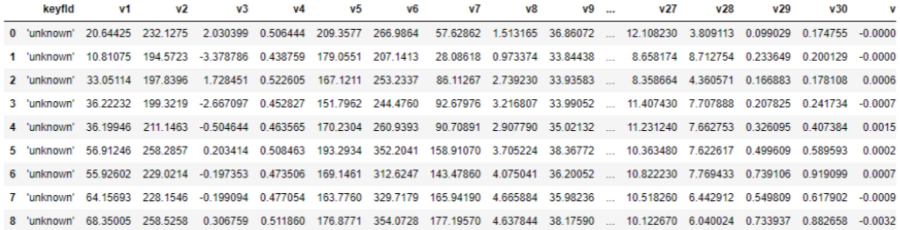
After acquiring IEMOCAP database's license agreement, the replication team was able to collect the author's IEMOCAP pre-processed dataset via an application link[5].

Figures 1 and 2 show snippets of processed MFCC and prosodic features of audio data, respectively.



	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

Figure 1: Extracted MFCC Features



	keyfld	v1	v2	v3	v4	v5	v6	v7	v8	v9	...	v27	v28	v29	v30	v
0	'unknown'	20.64425	232.1275	2.030399	0.506444	209.3577	266.9864	57.62862	1.513165	36.86072	...	12.108230	3.809113	0.099029	0.174755	-0.0000
1	'unknown'	10.81075	194.5723	-3.378786	0.438759	179.0551	207.1413	28.08618	0.973374	33.84438	...	8.658174	8.712754	0.233649	0.200129	-0.0000
2	'unknown'	33.05114	197.8396	1.728451	0.522605	167.1211	253.2337	86.11267	2.739230	33.93583	...	8.358664	4.360571	0.166883	0.178108	0.0006
3	'unknown'	36.22232	199.3219	-2.667097	0.452827	151.7962	244.4760	92.67976	3.216807	33.99052	...	11.407430	7.707888	0.207825	0.241734	-0.0007
4	'unknown'	36.19946	211.1463	-0.504644	0.463565	170.2304	260.9393	90.70891	2.907790	35.02132	...	11.231240	7.662753	0.326095	0.407384	0.0015
5	'unknown'	56.91246	258.2857	0.203414	0.508463	193.2934	352.2041	158.91070	3.705224	38.36772	...	10.363480	7.622617	0.499609	0.589593	0.0002
6	'unknown'	55.92602	229.0214	-0.197353	0.473506	169.1461	312.6247	143.47860	4.075041	36.20052	...	10.822230	7.769433	0.739106	0.919099	0.0007
7	'unknown'	64.15693	228.1546	-0.199094	0.477054	163.7760	329.7179	165.94190	4.665884	35.98236	...	10.518260	6.442912	0.549809	0.617902	-0.0009
8	'unknown'	68.35005	258.5258	0.306759	0.511860	176.8771	354.0728	177.19570	4.637844	38.17590	...	10.122670	6.040024	0.733937	0.882658	-0.0032

Figure 2: Extracted Prosodic Features

3 Replication of Original Work

Replications are done to validate the findings of the original work. To execute the replication successfully, first, the replicators must understand the goal of the research. The research proposed a novel deep dual recurrent model, which simultaneously utilise the text data and audio signals to analyze speech data. Original and pre-processed data,

[2] Papers that cite the research paper in Google Scholar:
`https://scholar.google.com.au/scholar?cites=16749771682156267984&as_sdt=2005&sciodt=0,5&hl=en`

[3] Author's github repository: `https://github.com/david-yoon/multimodal-speech-emotion`

[4] IEMOCAP database: `https://sail.usc.edu/iemocap/`

[5] Application to download IEMOCAP pre-processed dataset (SLT-18):
`https://docs.google.com/forms/d/e/1FAIpQLSdSrvnpoxltYnjq_QcFCl6GXfEco8G5JNEgpFuo56VhPM9Zbg/viewform`

tools, and guides were made available by the authors of the original work. The goal of the group is to successfully replicate the proposed model using the original data, then extend the replication to the group's new dataset. In order to do so, the structure of the data used in the original work was carefully studied and reviewed. Familiarization with the approach used in the paper as well as the environment the researchers used, ability to weigh the resources needed, and identification of potential roadblocks in the replication process are also aspects to focus on. The group re-evaluated the overall performance of the proposed model, based on the average of multiple training and evaluations, similar to the evaluation done in the paper.

3.1 Original Work Replication Issues

The codes were written in Python 2, for both Python script and notebooks. Linux commands are needed, to be able to run the Python scripts, and some were used inside the Python notebooks. TensorFlow 1.4.0, Scikit Learn 0.20.0, and Python 2.0 compiler were used in the research paper. With this, the group considered replicating the research using Linux Virtual Machine or Google Colaboratory. Listed are the issues faced by the group throughout the replication process.

Pre-processing of the Original Data

- The files sizes of original data are too big.
- Previous versions of Python, and other Linux commands are needed for the replication.
- Installation of openSMILE toolkit, and familiarization.
- Different output format of the new version of the openSMILE toolkit in emotion feature extraction.
- Some command line arguments do not work in Google Colaboratory notebook.

Model Execution using the Original Data

- Requires previous versions of Python and TensorFlow.
- No pre-trained model available.
- Model execution takes hours to train and evaluate.

3.2 Resolutions to Issues

The main issue that the group had for both pre-processing of the original data and model execution are the Python compiler version and tools versions. Installing the previous version of the compiler and tools needed are easier in the Virtual Machine. But the original dataset is composed of audio files, transcriptions, and label, which require large memory space and cannot be handled by the Virtual Machine due to the memory limit. Therefore, the group decided to use Google Colaboratory.

The current version of Python in Google Colaboratory is 3.6.9. To be able to run the pre-processing notebooks and scripts in Python 2.7, the group used a URL[6] that redirects to a Google Colaboratory notebook, which runs in a Python 2.7 compiler. Installation of previous versions TensorFlow and some Linux commands came after.

OpenSMILE toolkit is new to the group, and there are no references available for openSMILE installation in Google Colaboratory, which made the installation challenging. The group had to follow the installation guide from the openSMILE documentation for Linux, and had a few additional Linux command installation in Google Colaboratory required for openSMILE installation. Unlike in a Virtual Machine where tools need to be installed only once for the user to use them, Google Colaboratory needs to install the tools every time the user connects or uses the notebook. With this, the group added the step-by-step installation procedure of openSMILE and other Linux command installations at the beginning of the audio pre-processing notebook.

The openSMILE toolkit that the group had is version 2.3.0, a different version from what the authors used in the original work. In the emotion feature extraction, the pre-processing replication results has extra 42 lines

[6] Python 2 Google Colaboratory Notebook:
`https://colab.research.google.com/notebook#create=true&language=Python2`

before the expected extracted features data. This may be due to, version 2.3.0 producing different format of results, or the original researchers have not updated the pre-processing notebook. Since the group used the pre-processing notebooks the original researchers have provided in their GitHub repository, this issue was not anticipated. The group added a line of code in the audio pre-processing notebook, to fix these extra lines in the output file. All added codes are stored in the group's GitHub repository[7].

Aside from changing the path configurations in the original pre-processing notebooks, a few lines in the Python notebooks with Linux commands were altered. Due to some Linux commands not working, the `subprocess` library is used to execute the Linux commands. These lines of code were aligned with the syntax needed for the `subprocess()` commands to execute.

No pre-trained models are available for this research paper, and it was confirmed by one of the authors of the original work. The group then needed to re-execute the training and evaluation notebooks on the original data for the four models. Since the training and evaluation takes approximately 10 to 12 hours to compile per model, the group divided the model executions per group member.

After the successful execution of the models with the original dataset, the group have attained results close to the original work. Table 1 shows the overall performance of the replication and the original work, which has a small deviation for Text, Audio, and Multimodal models. A significant difference between the result of the original work and our replication in the Multimodal-Attention model is observed, which may be due to sub-optimal parameters, that is beyond the group's scope for research replication.

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

Table 1: Research and replication accuracy

4 Construction of New Data

Two approaches were considered when building the new dataset that will be tested on the models presented in the research. These two approaches are: creating a new one from scratch and using a dataset from a similar research. The development and pre-processing of the datasets are discussed in this section.

4.1 Approach 1: Creating Dataset from Scratch

In this approach, a good mix of audio and transcript that encompass the four emotions relevant to the research extension is desired. If the model were to be applied in a real-world scenario, audio and text have to be as close to reality as possible with overlapping dialogues and background noise. The group decided to download videos from the largest video sharing platform, YouTube, and extract their audio and captions using some of the techniques employed in the research. For the happy and sad emotions, romantic movie trailers were chosen since their themes predominantly involve combinations of cheerful and melancholic dialogues. For the angry emotion, a reality television episode that involves conflict and contains argumentative dialogues was selected. Neutral dialogues are very common in the videos and are interspersed between angry, happy and sad dialogues. The group ended up with using the three videos below:

- Dying Young Trailer[8]
- Trailers of Top 5 Romance Movies of All Time[9]
- UNTUCKED: Rupaul's Drag Race S09 E04[10]

[7] Replication GitHub Repository: `https://github.com/alyssa-raphaella/COMP8240_Project_J`

[8] `https://www.youtube.com/watch?v=A8p0w_Ec1NY`

[9] `https://www.youtube.com/watch?v=rLt-2dJRTxs`

[10] `https://www.youtube.com/watch?v=y0F_JE-dSxg`

Audio Extraction

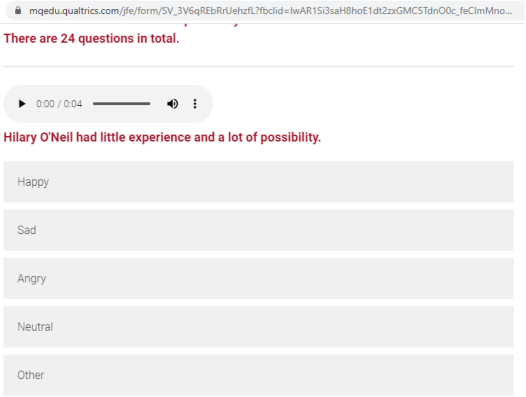
To download the audio from the three videos, Python and the `youtube_dl`[11] module were used. The output of the code is an mp3 file which was then converted into a wav file using the `sox`[12] module executed in the terminal. The wav file is the data format used in the research and subjected to further processing to extract features that will be used in the model. The audio files were cut to several files so that one file corresponds to a sentence uttered by a speaker. This was performed manually through audio editing software such as QuickTime.

Transcription Extraction

YouTube offers auto-generated captions through its own speech recognition algorithm or through user-generated captions provided by the content creators themselves. These captions are not directly available for download in the YouTube environment, so other methods for caption extraction had to be considered. The first extraction method was through Google Cloud Speech API[13], which was utilised in the research as well, however, upon inspecting the output, transcriptions look messy and distorted due to the background music of the videos. The second extraction method was through SaveSubs[14]. The site only needs the link to the YouTube video and outputs the captions into a text file. Although not 100% accurate, it produced better transcription than Google Cloud Speech API since YouTube captioning, somehow, is able to distinguish between the background music and dialogues. The group agreed to use the output of the second method and decided to manually rectify any errors in the text that may be due to speech overlaps, mispronunciations, accents or background noise. Similar to the audio files, the transcriptions were separated such that one line of the text file corresponds to a sentence uttered by a speaker. There are a total of 284 observations (audio with transcript) extracted from the three YouTube videos.

Emotion Labelling

Capturing human-annotated emotion labels for the audio and transcript extracted from the YouTube videos was the next step of the process. Amazon Mechanical Turk was considered but due to its cost, the group opted to carry-out the annotation process through surveys hosted on Qualtrics. The group constructed a number of surveys for each video, limiting the number of questions per survey to hold the attention of the annotators. The surveys contain the audio files and their corresponding transcripts and the multiple-choice options included Others so as not to limit the annotators to answer one of the four emotions relevant to the research. The list of survey links can be found in the Appendix section. A sample of the survey question in Qualtrics is shown in Figure 3.


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

There are 24 questions in total.

0:00 / 0:04

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

Happy

Sad

Angry

Neutral

Other

Figure 3: Sample Qualtrics survey question for emotion annotation

[11] `https://yt-dl.org`

[12] `http://sox.sourceforge.net`

[13] `https://cloud.google.com/speech-to-text/`

[14] `https://savesubs.com`

The emotion labels were collected from three annotators and the most common answer was used as the final label for the audio and its respective transcript. In case of no commonality in the answers, the group took it upon themselves to decide the final label of the observation which is still one of the answers by one of the annotators. The distribution of the emotions from the audio and transcript extracted from YouTube are shown in Figure 4.

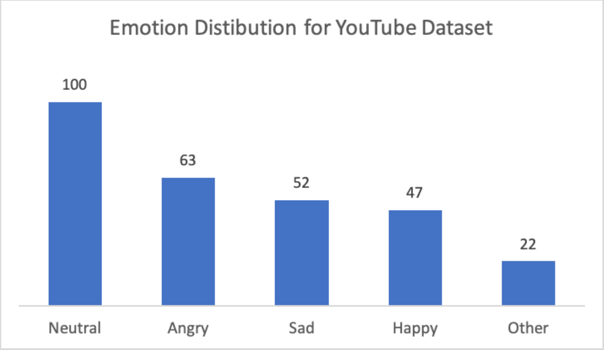


Figure 4: Emotion distribution for data collected on YouTube

The inter-annotator agreement was also calculated for the whole dataset and each individual video. Fleiss Kappa was utilised in the computation since the survey had answers from 3 different annotators. The results are displayed in Table 2.

Video	Fleiss Kappa
Dying Young Trailer	0.40
Trailers of Top 5 Romance Movies	0.45
UNTUCKED: Rupaul’s Drag Race	0.36
All videos	0.44

Table 2: Inter-annotator agreement for YouTube data

The inter-annotator agreement for the individual videos ranges from 0.36 to 0.45 while the overall inter-annotator agreement is 0.44. This level would be considered weak to moderate agreement and makes the annotations less accurate. As a final note, keys were created for each audio-transcript-label pair to link the observations together.

4.2 Approach 2: Using an Existing Dataset from a Similar Research

In this approach, the group did extensive research on previous studies or experiments that also tackled on a similar objective of predicting emotions using audio data. Most of the studies found also used the IEMOCAP dataset. An article from Towards Data Science titled Identifying Emotions from Voice using Transfer Learning[2] looked promising and used two audio datasets: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[15] [3]and Toronto Emotional Speech Set (TESS)[16] [4]. Both datasets were recorded by voice actors in a controlled environment, i.e., minimal background noise. The RAVDESS dataset, however, only contains 2 transcripts that is uttered by 24 different actors portraying different emotions. Hence, the TESS data was picked for this approach.

[15] https://zenodo.org/record/1188976#.X6P26ZMzamk

[16] https://tspace.library.utoronto.ca/handle/1807/24487

TESS Dataset Description

This dataset was collected by the University of Toronto, Psychology Department in 2010. There were 200 target words spoken by two actresses (aged 26 and 64 years) and recordings were made portraying each of seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The observations corresponding to the relevant emotions were then filtered for a total of 1,600 audio files (2 speakers, 200 transcripts and 4 emotions).

Transcription Extraction

The TESS data does not come with transcription, so the group had to extract them using Google Cloud Speech API which had 5.53% word error rate. It was expected to have a very high speech-to-text conversion accuracy since the audio files were recorded in a controlled environment.

Emotion Labels

The files are structured in a such a way that emotions can easily be extracted from each audio file. The amount of audio files are equally distributed among the four emotions with 400 observations each.

Similar to YouTube data, keys were created for each audio-transcript-label pair to link each observation in the TESS data.

4.3 Data Processing

Since the newly constructed dataset from YouTube and TESS come in a new format, it had to be processed to fit the models input parameters. The first step was to replicate the processing of the raw IEMOCAP data to get an idea of what type of inputs are needed by the model. This was performed on both audio and text data. It took a big chunk of the group's time since they had to deal with Python package conflicts, e.g., some NLTK commands are unable to load properly in Python 2 environment. When all the issues were resolved, the processing pipeline in the research was applied to the newly constructed datasets.

Text Processing

Two versions of text processing was performed. The first one was similar to the processing in the research which was very minimal and typical of processing in text analysis. It only includes transforming the transcripts to lower case and tokenizing them to words. The second version included additional text processing, which includes removing non-alphanumeric characters and stop words, and retaining relevant keywords only, i.e., nouns, verbs, adjectives and adverbs. Both text files were filtered for transcriptions that have emotion labels in the four emotion categories only and converted into a numpy array. There were 262 transcriptions left on YouTube data.

Audio Processing

To process the audio data and extract the relevant features, the research used the openSMILE toolkit. This was the biggest challenge in processing the audio data since the team had to quickly familiarise themselves with the platform. Performing the installation and processing in Virtual Machine did not go well due to storage constraints. It was agreed to install the toolkit on Google Colaboratory as well as the local machines, as a backup.

The audio feature extraction derived MFCC and prosodic features. The MFCC feature set contains 39 features: 12 MFCC parameters (1-12) from the 26 Mel-frequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients. The prosodic feature set contains 35 features such as F0 frequency, the voicing probability, and the loudness contours. These processing are both applied to YouTube and TESS datasets.

Similar filtering is done afterwards to retain the MFCC and prosodic features of observations that have the relevant emotion labels.

5 Results of New Data

The results of extending the research on YouTube and TESS datasets are shown side by side with the results published in the paper and the replication on the original data. Table 1 shows the retraining of the original data and the published results. Table 3, on the other hand, shows the accuracy results of the YouTube data alongside the results from the TESS data.

Model	YouTube Data	TESS Data
Text Only	32.7%	25.0%
Audio Only	23.5%	30.6%
Multimodal	27.3%	34.9%
Multimodal-Attention	23.9%	30.8%

Table 3: New data replication accuracy

As clearly shown, the evaluation on the new data resulted in accuracies way below the published results across the board. These results are best shown in the barchart in Figure 5.

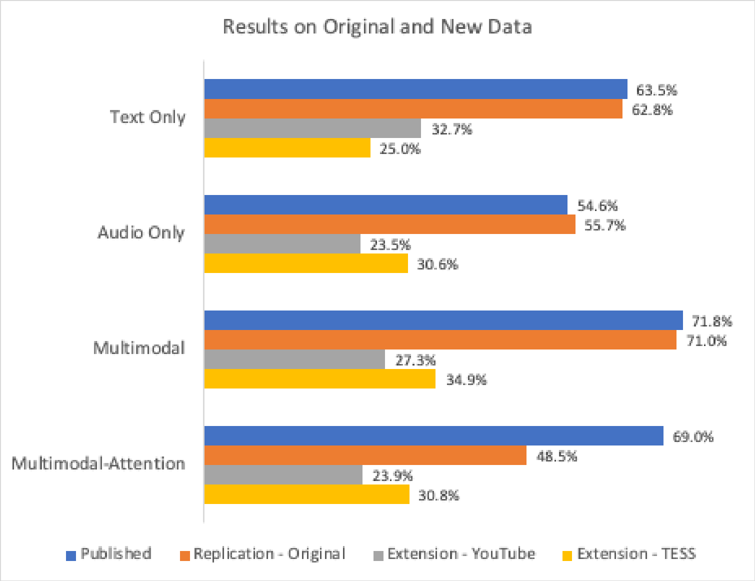


Figure 5: Replication on original data and extension on new data

The blue bar shows the benchmark published result in each of the models. The orange bar is the result of the replication on the original IEMOCAP data. All results except MDREA seems consistent with the published results.

The grey bar represents the results from the YouTube data whereas the yellow bar are for the TESS data. In both cases, the models barely achieved 30% accuracy.

The accuracy results achieved during the extension of the research were short of the published results. To the naked eye, this may ring bells of concern. Drilling further down, however, it should be recognised that there is a massive difference between the data used in the original training versus that used in this replication project. The techniques used here are completely different and the project is in fact an extension of the original model. For a maiden run of such new methodologies, this is in fact a good start and a good result at that. Further fine-tuning will improve on these results.

6 Summary and Reflections

Overall, this replication and extension project is deemed to have been successful given the limitations during the generation of the new data. The scores are in no doubt below the published results. However, the data used in the replication were very rudimentary and were not produced with the same level of sophistication as the original data. The low results here are just the initial results. It is still early days. Granted extra time, the replication team are confident that the techniques on generating new data can still be fine-tuned to hopefully end up with a higher accuracy result close to the published results. In addition, other factors may have contributed to the differences and may need to be reconsidered.

Absence of pre-trained model

The source codes were made available but the pre-trained model was missing hence the need to execute the training phase the models. It would have been more efficient had the model been available for at least three reasons - (1) elimination of the training time, (2) minimised chance of an introduced error and (3) established weights means more consistent accuracy.

By rerunning the training source codes, it costed the replication team about 10 to 12 additional hours. It also forced the team to rethink how to save the pre-trained model. In running the MDREA model, there was also an unidentified discrepancy in the accuracy figure a possible example of introduced error somewhere.

OpenSMILE configurations tuned to IEMOCAP data

OpenSMILE gives the users opportunity to create their own configuration files, and use it in extracting features with openSMILE. In this research, modified configuration files were created and used by the authors of the original work in extracting audio and emotion features from the IEMOCAP audio data. These modified configuration files may be tuned in IEMOCAP audio data, which might be the cause of the big difference in accuracy of the original data results and new data results. The configuration that the authors made, can perform good feature extractions on an audio recorded in a controlled environment like the IEMOCAP and TESS data, and not that well on audio with noisy background like the YouTube data.

Effect of TESS dataset variability

It was observed that in the TESS dataset, the transcripts have been shared between different emotions. The same transcripts have been spoken in different intonations which may have caused discrepancies in the prediction of labels especially those models utilising text data.

Effect of background music

YouTube videos were used in this project. These spelled two issues. Firstly, annotators may find that the tone of the background music affected how the emotions were labelled.

A positive word spoken with a sad music tone made annotators label them as sad.

Distortion of the audio sound is another issue. The original IEMOCAP dataset were collected in a controlled environment with potentially better acoustics. The YouTube videos in comparison have some noisy background.

Annotation improvements

The annotation surveys have received only three responses. A higher number of annotators would have improved the annotation accuracy. A high response will require a utility to pick the correct answer. One transcript, for instance, has been annotated as neutral, angry, and sad by different annotators. Which emotion will apply in these cases? A possible algorithm fix could be to pick the modal response. Say for instance that a transcript has 20 responses with 5 neutral, 14 angry and 1 sad, pandas function `db.mode()`, can label that transcript as angry. Additionally, Fleiss Kappa has been used to indicate inter-annotator agreement. With only three responses, the Fleiss Kappa resulted in a score of 0.44. This suggests a fair to moderate agreement [5] between the annotators which may denote inaccurate annotations and hence might explain the low accuracy score achieved in the YouTube dataset across all models.

References

[1] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. *Multimodal Speech Emotion Recognition Using Audio and Text*. 2018. arXiv: 1810.04635 [cs.CL].

[2] Shaan Shah. *Identifying Emotions from Voice using Transfer Learning*. https://towardsdatascience.com/detecting-emotions-from-voice-clips-f1f7cc5d4827. Accessed: 06-11-2020.

[3] Steven R. Livingstone and Frank A. Russo. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Version 1.0.0. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. Zenodo, Apr. 2018. DOI: 10.5281/zenodo.1188976. URL: https://doi.org/10.5281/zenodo.1188976.

[4] M. Kathleen Pichora-Fuller and Kate Dupuis. *Toronto emotional speech set (TESS)*. Version DRAFT VERSION. Scholars Portal Dataverse, 2020. DOI: 10.5683/SP2/E8H2MF. URL: https://doi.org/10.5683/SP2/E8H2MF.

[5] Mary McHugh. “Interrater reliability: The kappa statistic”. In: *Biochemia medica : asopis Hrvatskoga drutva medicinskih biokemiara / HDMB* 22 (Oct. 2012), pp. 276–82. DOI: 10.11613/BM.2012.031.

Appendix

Complete list of surveys used for YouTube data emotion annotation.

Dying Young Trailer

- https://mqedu.qualtrics.com/jfe/form/SV_3V6qREbRrUehzfL

Trailers of Top 5 romance movies of all time

- https://mqedu.qualtrics.com/jfe/form/SV_aWfIqIE47zNqQiF
- https://mqedu.qualtrics.com/jfe/form/SV_9zs3qKmWTmEXIaN
- https://mqedu.qualtrics.com/jfe/form/SV_2mJfhZUqAeBOHyd
- https://mqedu.qualtrics.com/jfe/form/SV_5gS1alMioRNID5j
- https://mqedu.qualtrics.com/jfe/form/SV_4YHTSBAAnPTWmeV
- https://mqedu.qualtrics.com/jfe/form/SV_6ETL75m3gZ8GneB
- https://mqedu.qualtrics.com/jfe/form/SV_4UsoeTpDaejmyvH
- https://mqedu.qualtrics.com/jfe/form/SV_0ddSfHUv29ljs3P
- https://mqedu.qualtrics.com/jfe/form/SV_egtQkblhurwcGoZ
- https://mqedu.qualtrics.com/jfe/form/SV_1Gp9SQeOXHhe2ot
- https://mqedu.qualtrics.com/jfe/form/SV_bwQNmDokIGzulQ9
- https://mqedu.qualtrics.com/jfe/form/SV_6hc6nkmrPOQYgyp
- https://mqedu.qualtrics.com/jfe/form/SV_diBPgwl6G0FWSbP
- https://mqedu.qualtrics.com/jfe/form/SV_54quaKRcyY3hBrL
- https://mqedu.qualtrics.com/jfe/form/SV_5dzcJEMHMPMpcuF
- https://mqedu.qualtrics.com/jfe/form/SV_1RAZxwNgXhj5QXP
- https://mqedu.qualtrics.com/jfe/form/SV_ekXULT5y7ucumQR

UNTUCKED: Rupaul’s Drag Race S09 E04

- https://mqedu.qualtrics.com/jfe/form/SV_82F4MKD2h4Chc8Z
- https://mqedu.qualtrics.com/jfe/form/SV_aicSb3q30AuU8YJ
- https://mqedu.qualtrics.com/jfe/form/SV_2cynAKMsqwGvQRD