\documentclass{article}

\usepackage[utf8]{inputenc}

\usepackage{geometry}

\usepackage{graphicx}

\usepackage{hyperref}

\usepackage{listings}

\usepackage{color}

\usepackage{pgfplotstable}

\graphicspath{ {./images/} }

\lstset{

language=R,

basicstyle=\small\sffamily,

numbers=left,

numberstyle=\tiny,

frame=tb,

tabsize=4,

columns=fixed,

showstringspaces=false,

showtabs=false,

keepspaces,

commentstyle=\color{red},

keywordstyle=\color{blue}

}

\title{MULTIMODAL SPEECH EMOTION RECOGNITION USING AUDIO AND TEXT \\

\large A MODEL REPLICATION PROJECT}

% \title{A REPLICAT}

\author{Jan Arvin Lapuz, Alyssa Lim, Le Van Nguyen, Ramil Zabala}

% \date{September 2020}

\begin{document}

\maketitle

% \section\*{Abstract}

% \large

\\

% The segmentation marker used was a string of ten equal sign markers (==========) which separates the documents into paragraphs with the same topic.

\\

\\

\section\*{Brief Description of the Source Paper and Justification}

\large

\hspace{5mm}Extensive studies have been done on creating models that will classify emotions through speech audio or through texts. But never has a study been conducted on making a more comprehensive prediction by combining these two.

\\

\\

This model includes dual recurrent neural network (RNNs) that will combine both audio and text to predict the emotion into four categories - angry, happy, sad and neutral. The paper outperformed earlier prediction models with accuracy results ranging from 68.8\% to 71.8\%.

\\

\\

The paper was presented during the 2018 IEEE Spoken Language Technology Workshop (SLT) and it ranked 14 in Computational Linguistics Category on Google Scholar.

\\

\\

\section\*{Description of Original Dataset}

\large

\\

\hspace{5mm}The research team openly provided both the source codes and the full set of data used during the training of the model.

\\

\\

\\

The model was trained and evaluated using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. This dataset was generated by interactions between 2 actors over 5 sessions thus creating 10 unique speakers. The dataset contains 5531 utterances.

\\

\\

To extract the audio and the prosody components, the .wav files have been fed into Open Smile 2.3.0. This will extract the required MFCC parameters that will be used for the audio and the prosody component which contains the emotion data.

\\

\\

The same .wav files were then fed into Google Cloud Speech API to retrieve the transcript. NLTK 3.3 was then used to tokenize the text

\\

\\

\textbf{Wav file}

.... insert snippet of the wav file here ...

\\

\\

\textbf{Text file}

\\

.... insert snippet of the text file here ...

\\

% • Description of original dataset: This only applies to novel projects, not the default project. Explain the format of the data, and show a snippet of it (in e.g. a figure). You will typically give a URL link to the dataset in a footnote.

\\

\\

\section\*{Model Structure}

\large

\\

\includegraphics[width=150mm]{images/Model.png}

\\

\\

\section\*{Replication of Original Work}

\large

\hspace{5mm}The source codes were available, however, the pre-trained model was not. Group J needed to retrain and evaluate using the raw source codes.

\\

The model was trained in a Python 2.7 environment. With recent Python version being 3.8, Python 2.7 was initially installed on VM but with space limitations, it was subsequently ran on Colab.

\\

The neural networks were trained on Tensorflow 1.4.0 and reports were generated through Scikit Learn 0.20.0.

\\

\textit{... add discussion of replication here ....}

\\

\item Replication Results: \\

\begin{center}

\begin{tabular}{||c c c||}

\hline

Model & Accuracy & Research \\ [0.5ex]

\hline\hline

Text Only & 62.3\% & 63.5\%\\

\hline

Audio Only & 53.7\% & 54.6\%\\

\hline

Multimodal & 71.9\% & 71.8\%\\

\hline

Multimodal-Attention & 54.6\% & 69.0\%\\

\hline

\end{tabular}

\end{center}

\\

% • Replication of original work: This, and all the following sections, apply to both the novel projects and the default project. The replication concerns getting original code working on original data, and getting something like the original results. In some cases (e.g. if there's a mistake in the original work) it might not be possible to get the original results, but the project report should contain evidence of investigating this. This section should have some detail on the environment used to get the original code working, etc. You will typically give a URL link to your github repo or other repository where you have stored the scripts you used in this.

\\

\section\*{Construction of New Data}

\large

Two approaches were considered in producing the new data - (1) Using data from similar studies, and (2) Creating the data from scratch.

\\

$\textbf{Data from Similar Studies}$

\\

Two similar researches classifying audio into emotions have been found - RAVDESS and TESS.

For RAVDESS, they create 8 emotions whereas TESS focuses on 7 emotions.

\\

This paper will only use 4 of these emotions and hence, will only use a subset of the RAVDESS dataset.

\\

\textbf{Data from Scratch}

Three youtube videos with different emotions have been downloaded from Youtube - Dying Young, puppy video, and < ... >

The mp3 files have been converted to .wav format.

% which was subsequently fed into OpenSmile to extract the MFCC and prosody components.

\\

To capture the text, the youtube captions were downloaded using < ...> and saved into a text file. The text files have been pre-processed to the required format.

\\

To capture the emotion from the wav time, Quicktime was used to split the audio into chunks based on sentences.

\\

\\

A Qualtrics survey has been created to ask online annotators to classify the emotion conveyed by each of the sentences and the sound.

\\

\\

The survey questions attached a clip of the audio sound which can be heard by the annotators. The corresponding text is shown. Annotators were given four choices of emotions - happy, sad, angry, neutral and other (rest of the emotions). An example survey is shown below.

\\

\\

\includegraphics[width=120mm]{images/Qualtrics\_Survey.png}

\\

\\

The survey results have been collated and Cohen Kappa has been calculated to check annotator agreement. The Cohen Kappa resulted in 0.65 indicating a solid agreement between annotator responses. The index calculation is shown below.

\\

\\

\includegraphics[width=100mm]{images/Kappa\_Cohen.png}

\\ The sentence texts have been reformatted to a .csv format.

\\

The .wav file was fed into Open Smile 2.3.0 to extract the MFCC and prosody features in .csv format.

\\

\\

$\emph{MFCC Format}$

\\

The MFCC File contains 39 features which includes 12 MFCC parameters from the Melfrequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients. $^{[1]}$ A snippet of the MFCC dataset is shown below.

\\

\includegraphics[width=120mm]{images/MFCC File Layout.png}

\\

$\emph{Prosody File Format}$

\\

The prosodic features are composed

of 35 features, which include the F0 frequency, the voicing

probability, and the loudness contours.$^{[1]}$ A snippet of the prosody dataset is shown below.

\\

\\

\includegraphics[width=120mm]{images/Prosody File Structure.png}

\\

\\

<include Text file snippet here>

\\

% • Construction of new data: In this section, you will explain where you obtained your data from, and what processing was involved in preparing it as input for your chosen system. You should show snippets of your data (in e.g. a figure). You should have one subsection per new dataset. If your datasets required human annotation, explain how you did this, and provide a measure of agreement.

\section\*{Results of New Data}

\large

\\

The preprocessed data was then fed into the model.

<discuss the process that was undertaken here>

Results are shown below.

\\

\item Replication Results: \\

\begin{center}

\begin{tabular}{||c c c||}

\hline

Model & Accuracy & Research \\ [0.5ex]

\hline\hline

Text Only & 62.3\% & 63.5\%\\

\hline

Audio Only & 53.7\% & 54.6\%\\

\hline

Multimodal & 71.9\% & 71.8\%\\

\hline

Multimodal-Attention & 54.6\% & 69.0\%\\

\hline

\end{tabular}

\end{center}

\\

It appears that ...

\\

\\

% • Results on new data: In this section, you will describe the framework for your constructed datasets. If this is the same framework as for the original dataset (e.g. accuracy for a machine learning classification system), just briefly refer back to your earlier description of the source paper (for a novel project). If you use a different framework (e.g. using human evaluation rather than classification accuracy), explain the chosen evaluation and its motivation.

\section\*{Reflections}

\large

The source code was available but it does not include the pre-trained model. Hence the need to retrain the data to produce the results. It would have been more efficient had the model been available.

\\

\\

Other reflections ...

\\

\\

% • Any other reflections: The final section is optional, and can contain any other reflections -- the realisation that another approach to dataset construction would have been better, or that annotation is unexpectedly difficult for this task, etc.

% As noted in the overall project guidelines from the start of semester, the assessment will consider the extent to which the unit techniques and tools are incorporated into the project. These should be incorporated into appropriate sections above. Examples include the use of scripts to gather data from the Internet, carry out data analysis, and generate new data. The project does not have to incorporate every topic covered in the unit, but should manage several.

% \section\*{Bibliography}

% \large

% \bibliography{refs}

\begin{thebibliography}{9}

% \bibitem{latexcompanion}

% Michel Goossens, Frank Mittelbach, and Alexander Samarin.

% \textit{The \LaTeX\ Companion}.

% Addison-Wesley, Reading, Massachusetts, 1993.

% \bibitem{einstein}

% Albert Einstein.

% \textit{Zur Elektrodynamik bewegter K{\"o}rper}. (German)

% [\textit{On the electrodynamics of moving bodies}].

% Annalen der Physik, 322(10):891–921, 1905.

\bibitem{knuthwebsite}

Multimodal Speech Recognition Using Audio and Text,

\\\texttt{https://arxiv.org/abs/1810.04635}

\bibitem{knuthwebsite}

RAVDESS Dataset

\\\texttt{https://zenodo.org/record/3255102#.X4\_Ed9AzbIU}

\bibitem{knuthwebsite}

Raw Source Codes

\\\texttt{https://github.com/david-yoon/multimodal-speech-emotion}{https://github.com/david-yoon/multimodal-speech-emotion}

{}

\end{thebibliography}

\end{document}