TEAM 13

FINAL PROJECT

CPE 313

Advanced Machine Learning and Deep Learning

Dejoras, Dylan James   *Villamor, Kurt Russel*

NLP AND STT SYSTEMS STRUGGLE WITH SENTIMENT DETECTION DUE TO BIASES, NOISE, AND CHALLENGES IN PROCESSING SPONTANEOUS SPEECH AND LOW-RESOURCE LANGUAGES.

SOURCE:
HTTPS://WWW.RESEARCHGATE.NET/POST/WHAT_ARE_THE_CHALLENGES_RESEARCHERS_MAY_FACE_DURING_LINGUISTIC_TEXT_ANALYSIS_HOW_CAN_THESE_CHALLENGES_BE_OVERCOME

Speech-to-Text and Tonal Analysis for VOICE-BASED SENTIMENT ANALYSIS

using Small Language Modeling

Dejoras, Dylan James   *Villamor, Kurt Russel*

INTRODUCTION

Sentiment analysis helps industries improve user experience by using both text and voice data, with voice-based analysis capturing emotions through tone, pitch, and speech speed.

OBJECTIVES

- Compare sentiment classification using text-based and audio-based inputs.
- Evaluate practical performance of each input type in classification tasks.
- Explore deep learning models for audio-based and text-based sentiment recognition.

INTRODUCING LANGUAGE MODELS

Bidirectional Encoder Representations from Transformers

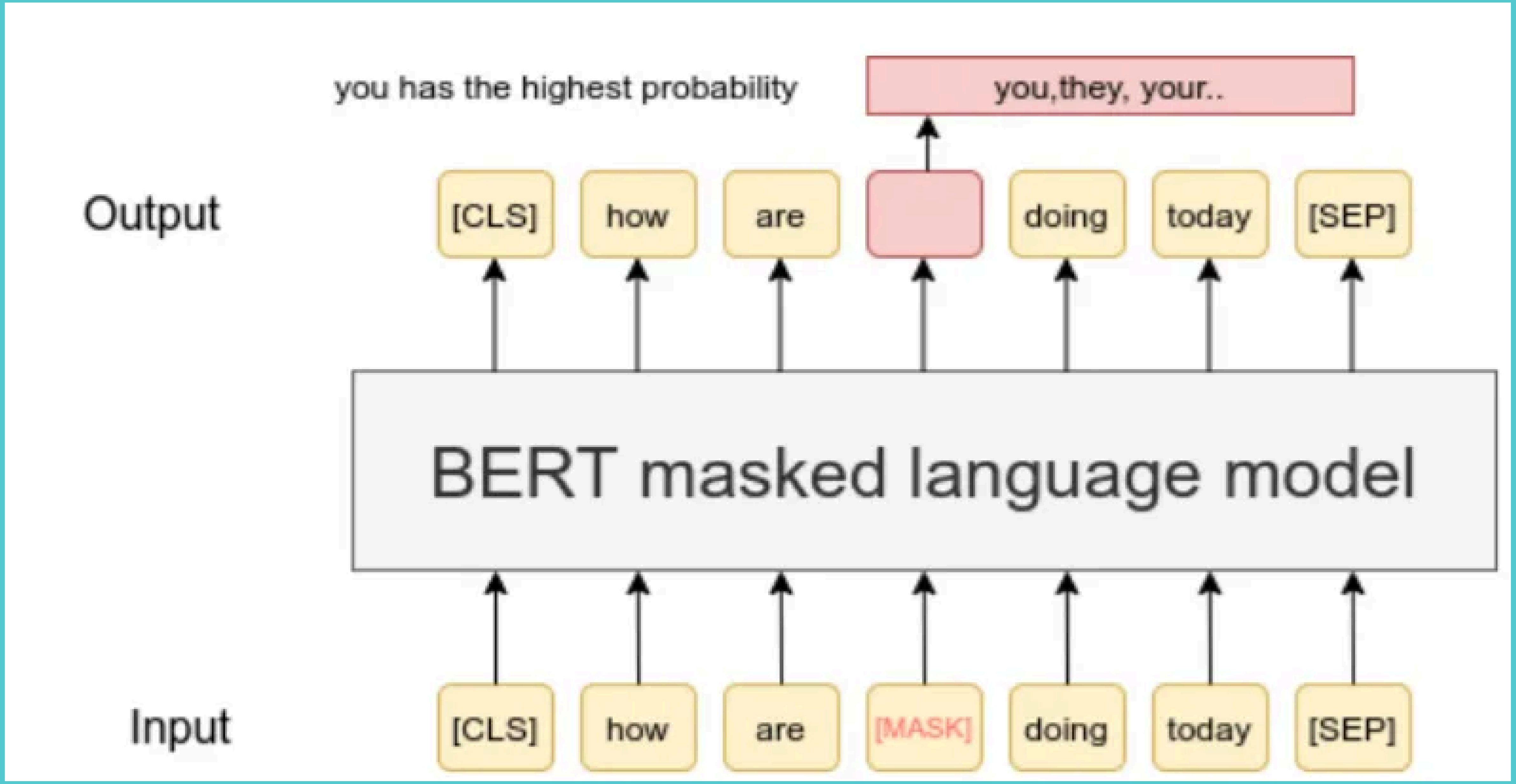
&

Efficiently Learning an Encoder that Classifies Token Replacements Accurately

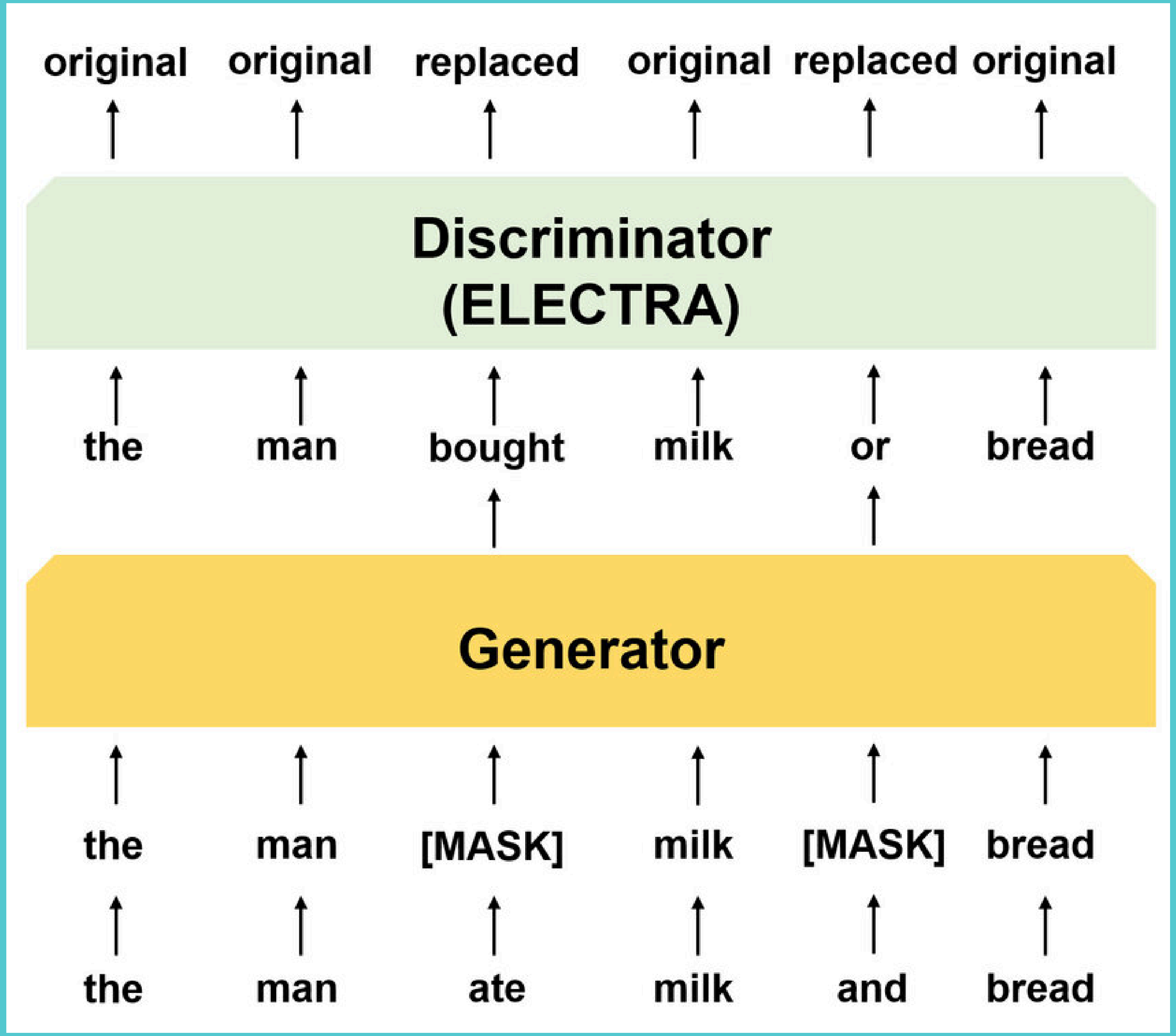
BERT

ELECTRA

BERT


you has the highest probability
you,they, your..
Output
[CLS]
how
are
doing
today
[SEP]
BERT masked language model
Input
[CLS]
how
are
[MASK]
doing
today
[SEP]

ELECTRA


original
original
replaced
original
replaced
original
Discriminator (ELECTRA)
the
man
bought
milk
or
bread
Generator
the
man
[MASK]
milk
[MASK]
bread
the
man
ate
milk
and
bread

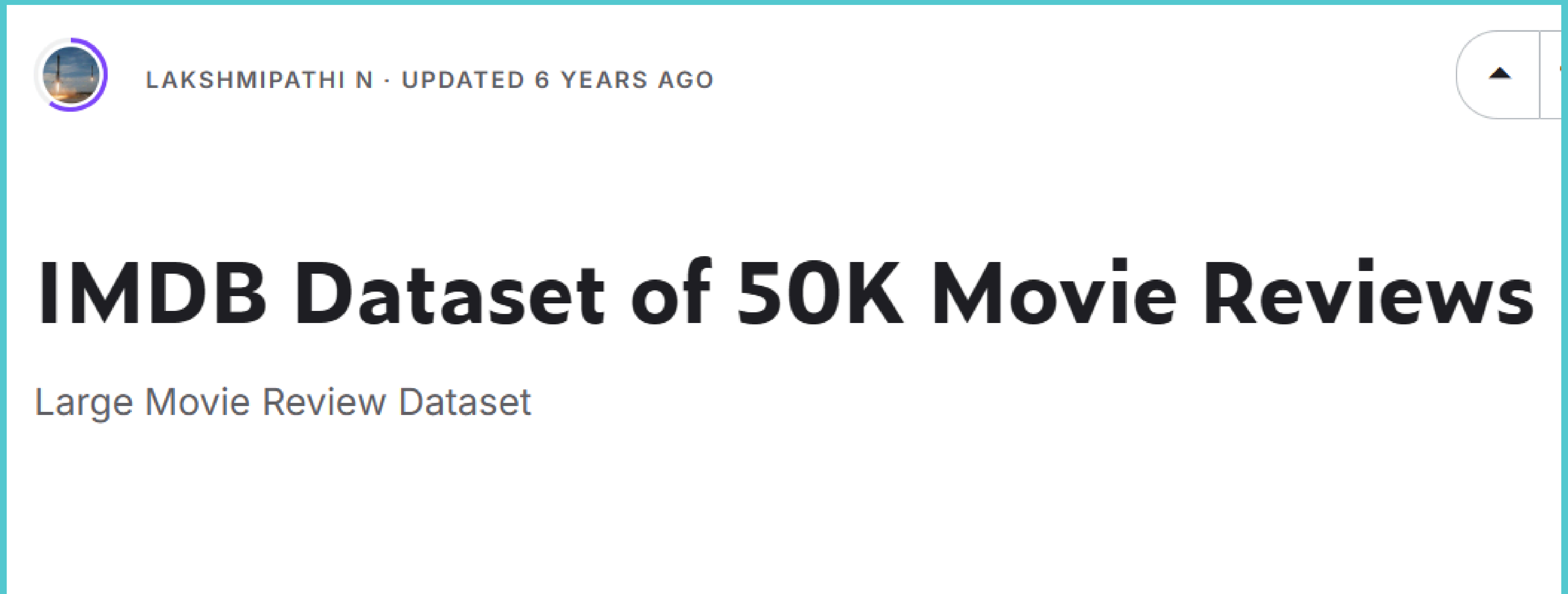
MODEL VARIANTS

RoBERTa

ELECTRA-Small

TinyBERT

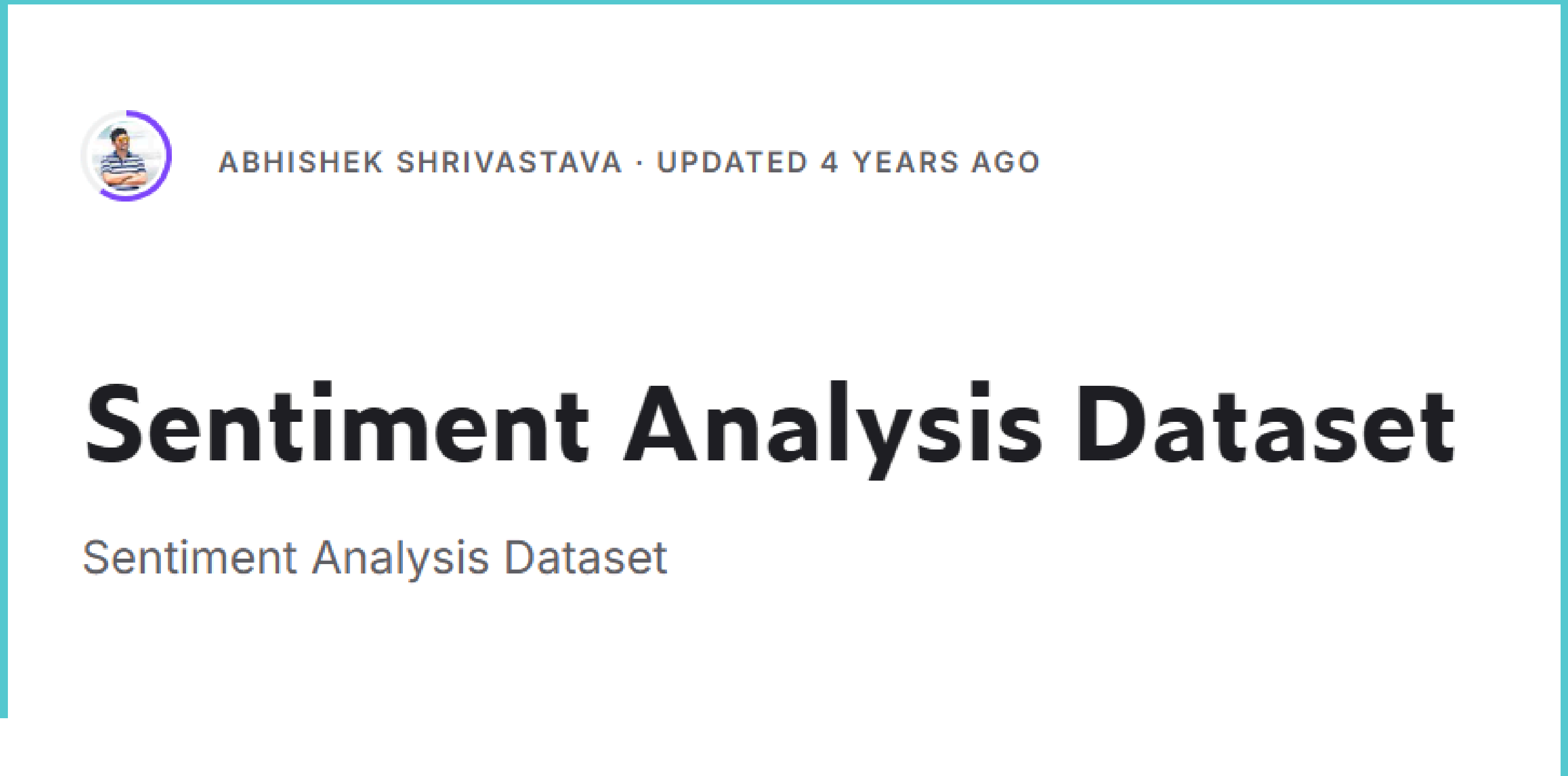
DATASETS



LAKSHMIPATHI N · UPDATED 6 YEARS AGO

IMDB Dataset of 50K Movie Reviews

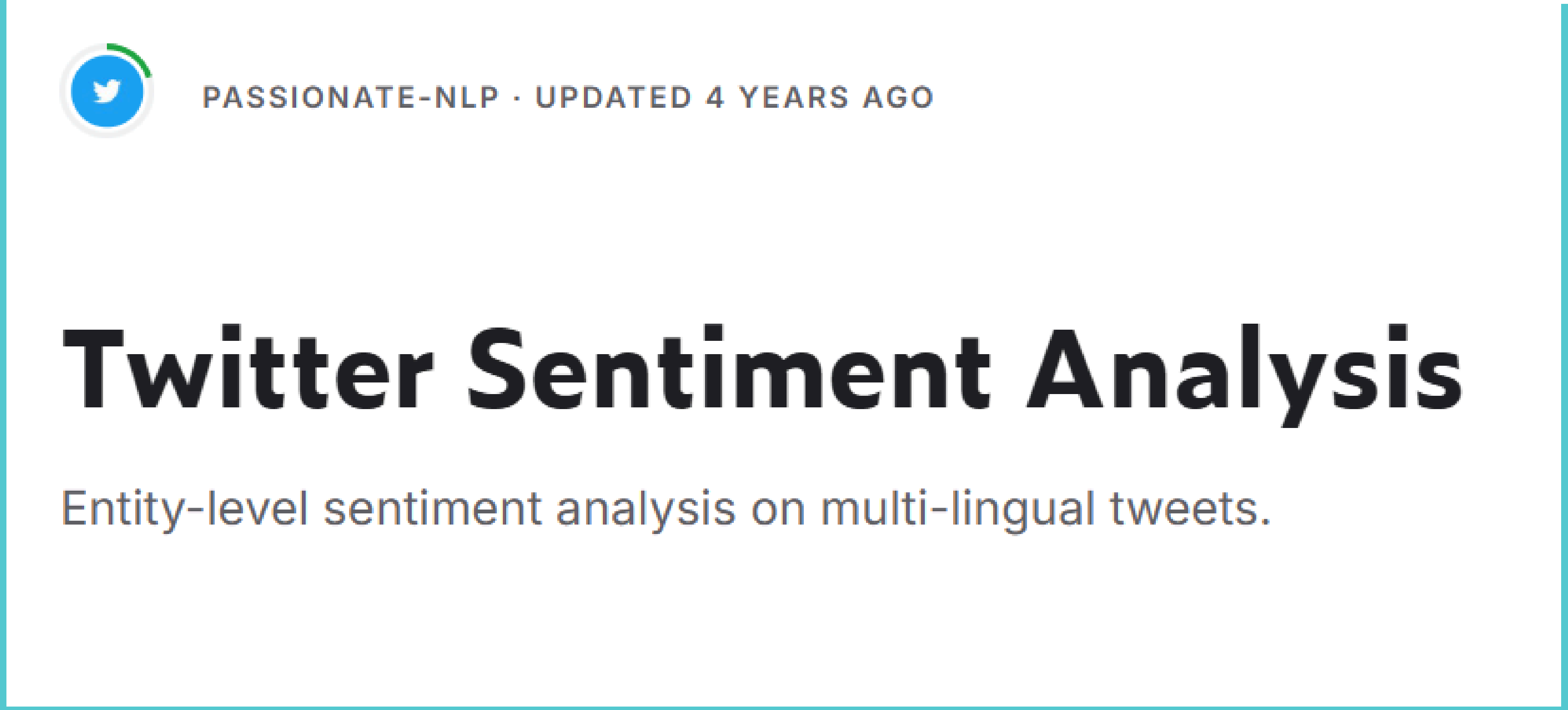
Large Movie Review Dataset



ABHISHEK SHRIVASTAVA · UPDATED 4 YEARS AGO

Sentiment Analysis Dataset

Sentiment Analysis Dataset

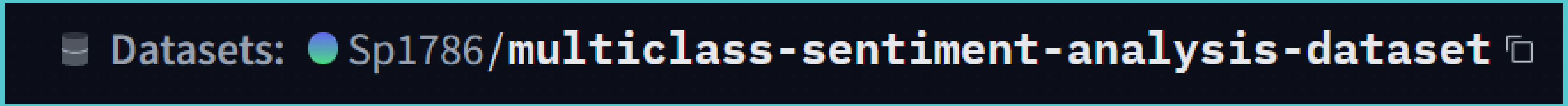


PASSIONATE-NLP · UPDATED 4 YEARS AGO

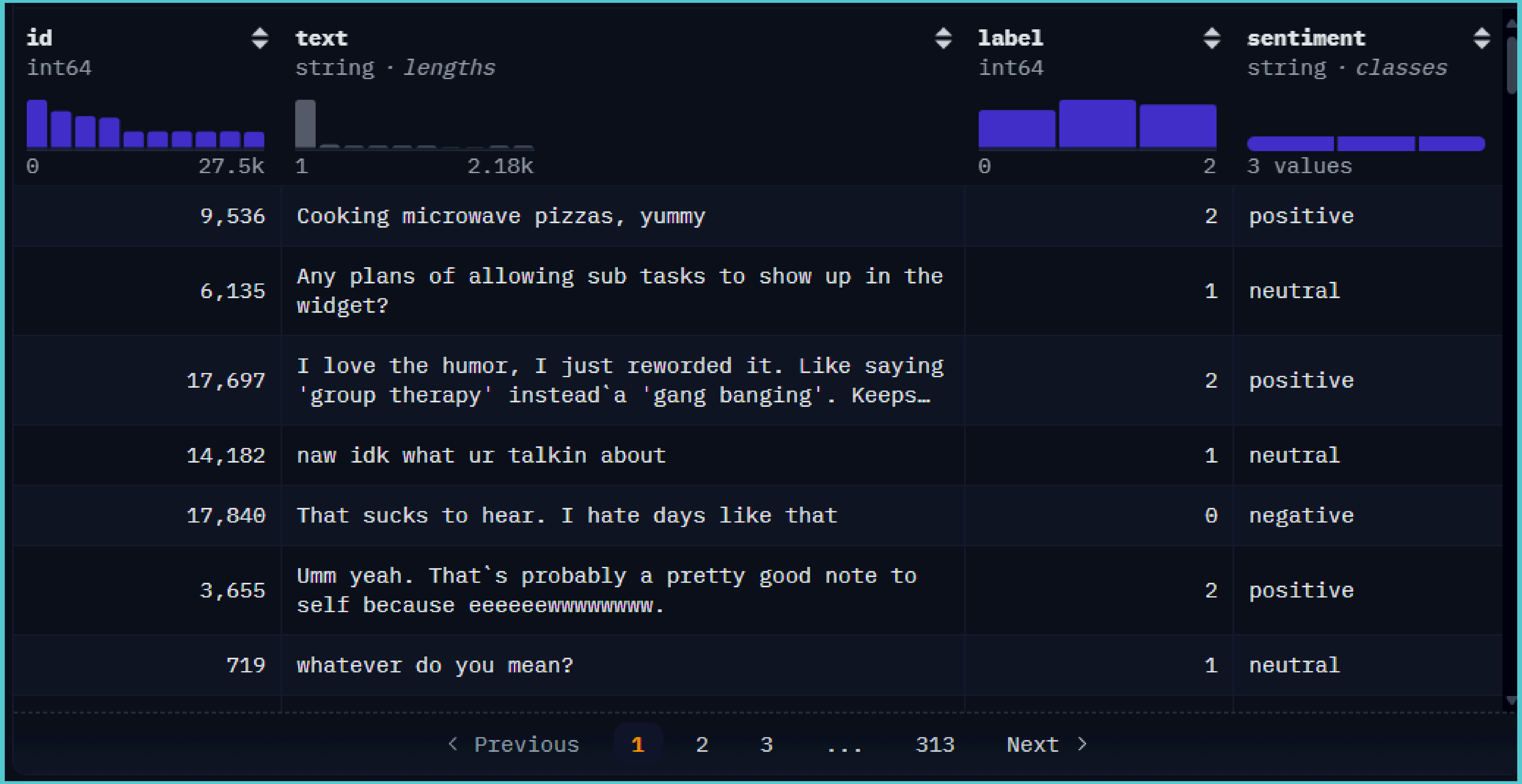
Twitter Sentiment Analysis

Entity-level sentiment analysis on multi-lingual tweets.

DATASETS



Datasets: Sp1786/**multiclass-sentiment-analysis-dataset**



id int64 0 27.5k	text string · *lengths* 1 2.18k	label int64 0 2	sentiment string · *classes* 3 values
9,536	Cooking microwave pizzas, yummy	2	positive
6,135	Any plans of allowing sub tasks to show up in the widget?	1	neutral
17,697	I love the humor, I just reworded it. Like saying 'group therapy' instead`a 'gang banging'. Keeps…	2	positive
14,182	naw idk what ur talkin about	1	neutral
17,840	That sucks to hear. I hate days like that	0	negative
3,655	Umm yeah. That`s probably a pretty good note to self because eeeeeewwwwwwww.	2	positive
719	whatever do you mean?	1	neutral

118,334 samples

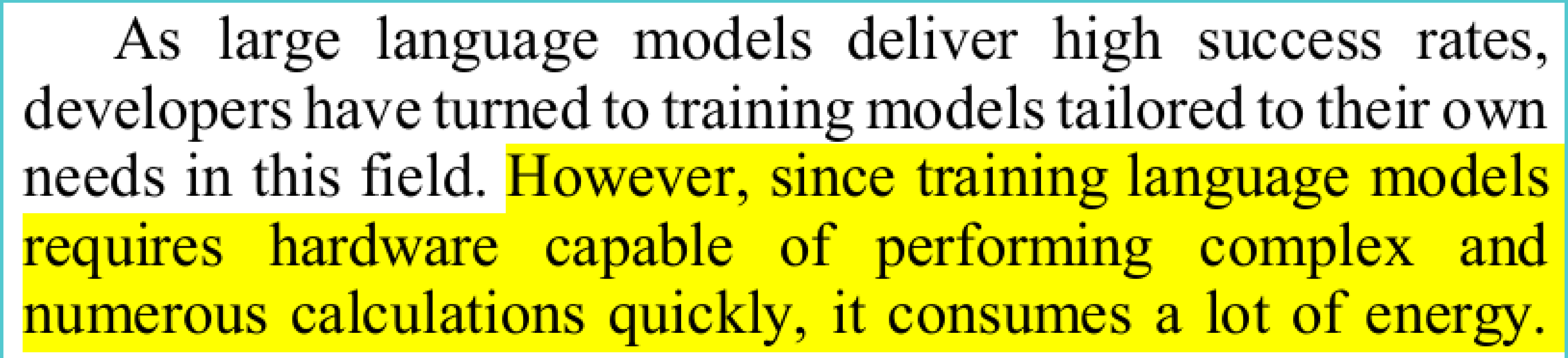
MODEL VARIANTS

	MobileBERT	TinyBERT	ELECTRA-Small
ACCURACY	89%	81%	89%
SIZE	96.5 mb	17.15 mb	53.01 mb
TRAINING TIME	9hrs. 20min.	1hrs. 47min.	2hrs. 4min.

MODEL VARIANTS

	MobileBERT	TinyBERT	ELECTRA-Small
ACCURACY	✓	✗	✓
SIZE	✓	✓	✓
TRAINING TIME	✗	✓	✓

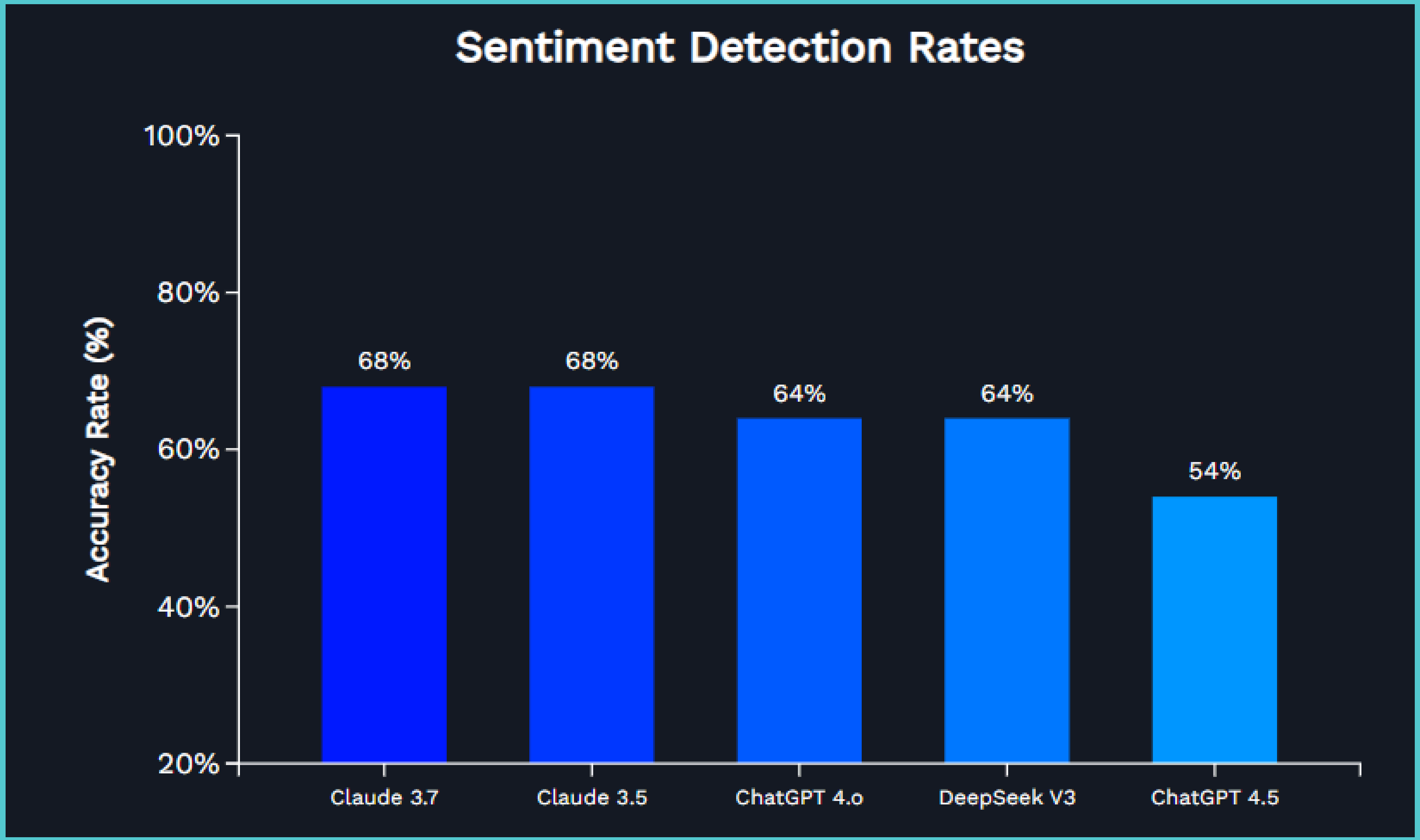
THE CHALLENGE OF LIMITED RESOURCES



As large language models deliver high success rates, developers have turned to training models tailored to their own needs in this field. However, since training language models requires hardware capable of performing complex and numerous calculations quickly, it consumes a lot of energy.

SOURCE: Z. ÖRPEK, B. TURAL AND Z. DESTAN, "THE LANGUAGE MODEL REVOLUTION: LLM AND SLM ANALYSIS," 2024 8TH INTERNATIONAL ARTIFICIAL INTELLIGENCE AND DATA PROCESSING SYMPOSIUM (IDAP), MALATYA, TURKIYE, 2024, PP. 1-4, DOI: 10.1109/IDAP64064.2024.10710677.

LARGE LANGUAGE MODELS IN SOME CASES



SOURCE: HTTPS://RESEARCH.AIMULTIPLE.COM/SENTIMENT-ANALYSIS-BENCHMARK/#BENCHMARK-DATASET-AND-METHODOLOGY

PROCESS

1

SPEECH AS

INPUT

2

TEXT TRANSCRIPT USING

SPEECH-TO-TEXT MODEL

3

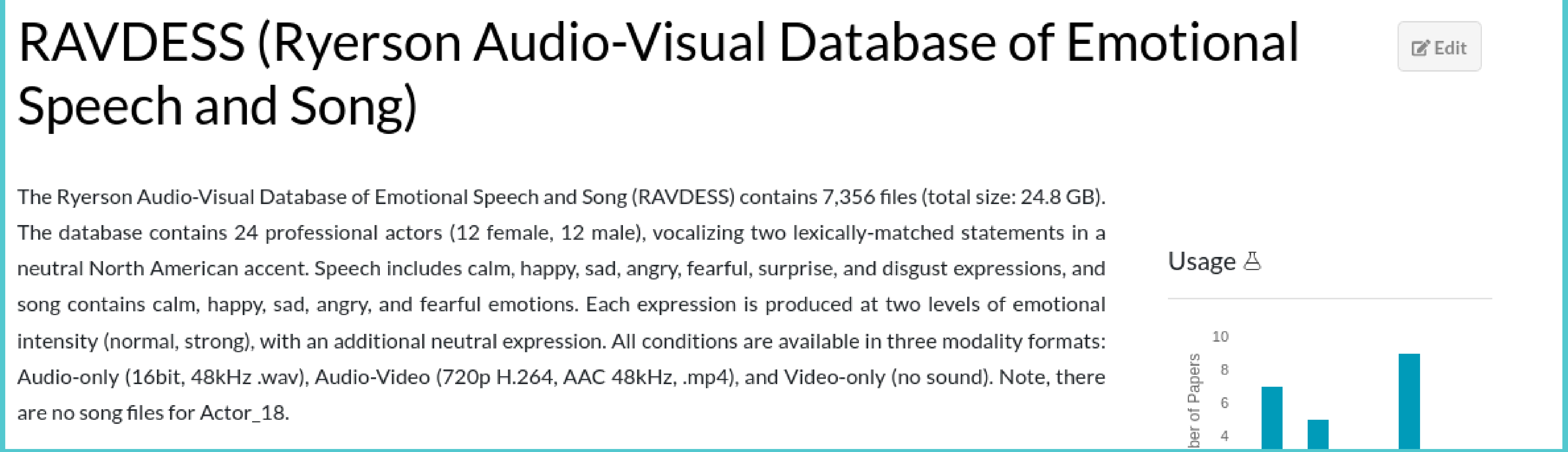
CLASSIFIED SENTIMENT

(POSITIVE, NEUTRAL, NEGATIVE)

VOICE-BASED

SENTIMENT ANALYSIS

DATASETS



RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7,356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Note, there are no song files for Actor_18.

RAVDESS

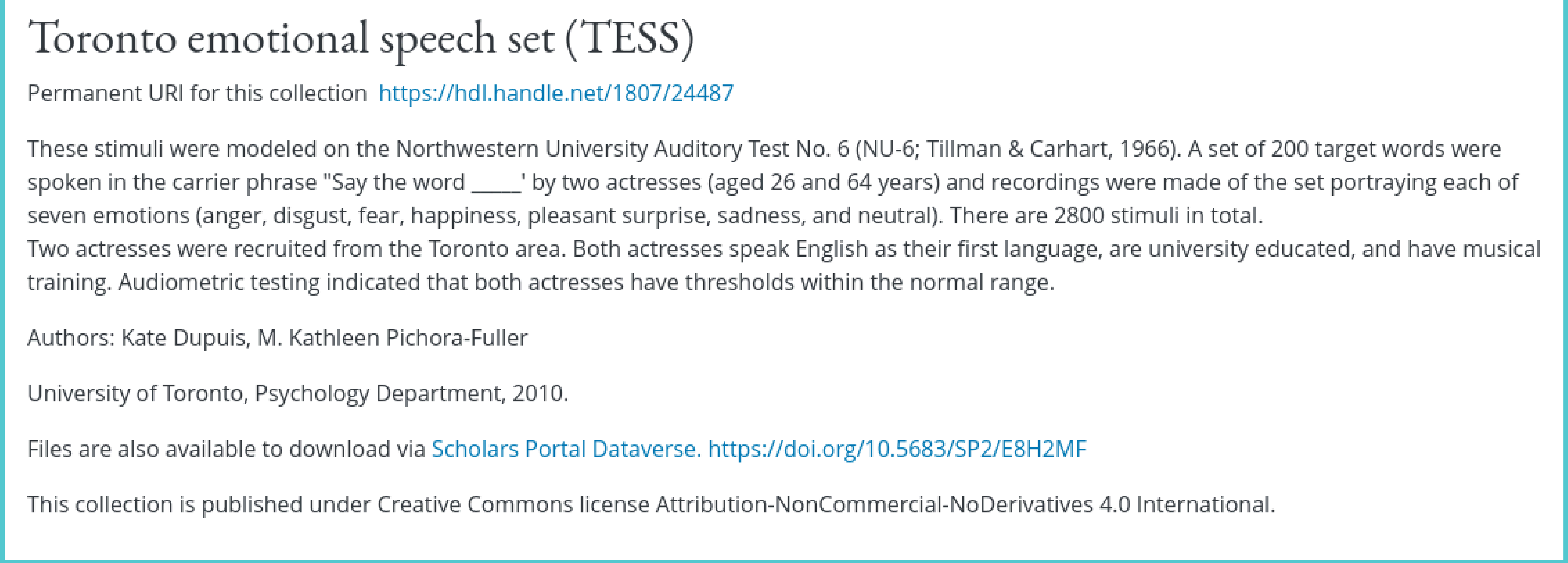
'NEUTRAL', 'CALM', 'HAPPY', 'SAD', "ANGRY', 'FEARFUL', "DISGUST', 'SURPRISED'

2304 SAMPLES (RESAMPLED)

TESS

'NEUTRAL', 'CALM', 'HAPPY', 'SAD', "ANGRY', 'FEARFUL', "DISGUST', 'SURPRISED'

4800 SAMPLES (RESAMPLED)



Toronto emotional speech set (TESS)

Permanent URI for this collection https://hdl.handle.net/1807/24487

These stimuli were modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966). A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total.
Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

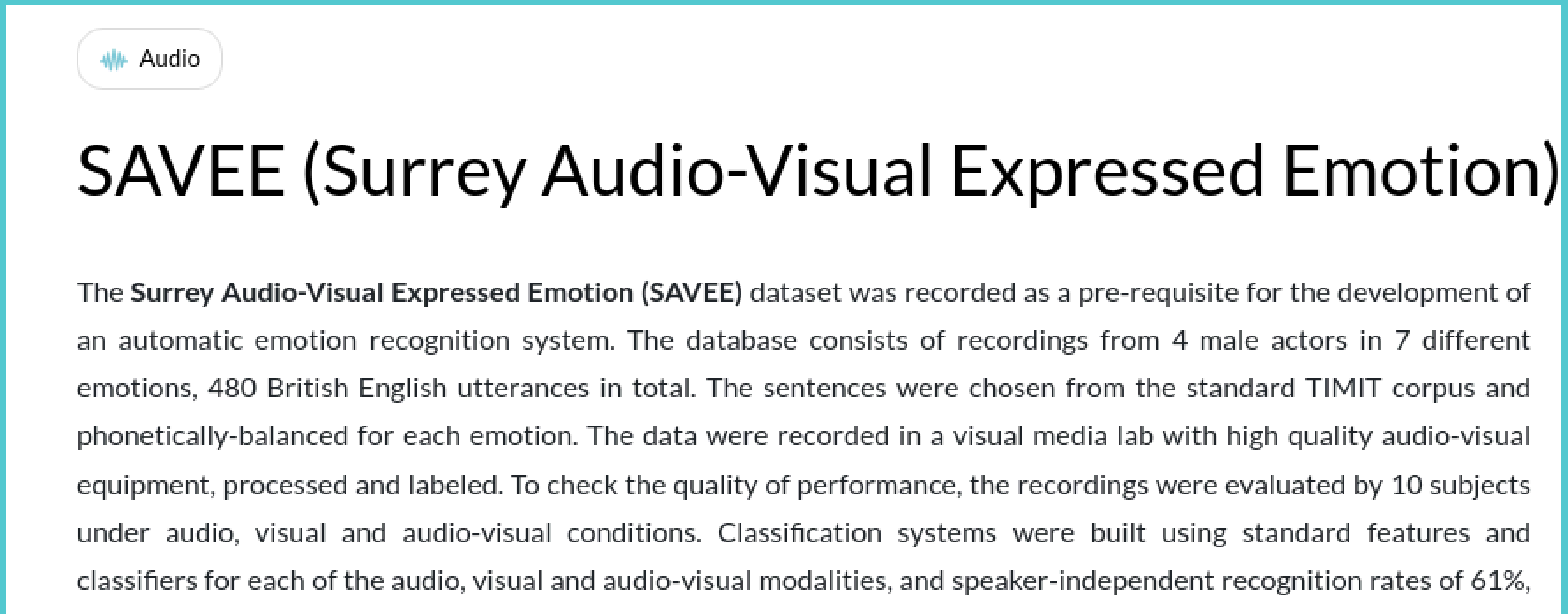
Authors: Kate Dupuis, M. Kathleen Pichora-Fuller

University of Toronto, Psychology Department, 2010.

Files are also available to download via Scholars Portal Dataverse. https://doi.org/10.5683/SP2/E8H2MF

This collection is published under Creative Commons license Attribution-NonCommercial-NoDerivatives 4.0 International.

DATASETS



Audio

SAVEE (Surrey Audio-Visual Expressed Emotion)

The **Surrey Audio-Visual Expressed Emotion (SAVEE)** dataset was recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. The data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled. To check the quality of performance, the recordings were evaluated by 10 subjects under audio, visual and audio-visual conditions. Classification systems were built using standard features and classifiers for each of the audio, visual and audio-visual modalities, and speaker-independent recognition rates of 61%, 65% and 84% achieved respectively.

SAVEE

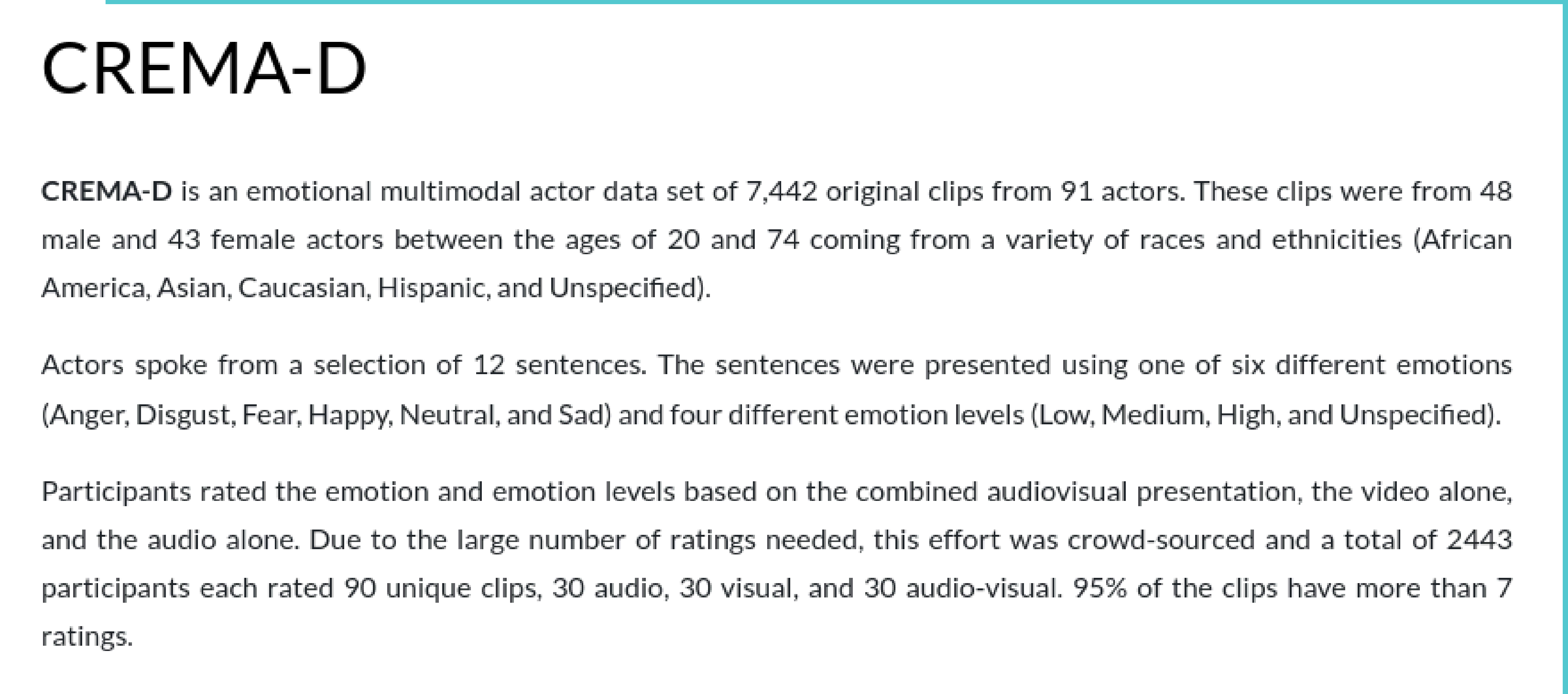
'NEUTRAL', 'CALM', 'HAPPY', 'SAD', "ANGRY', 'FEARFUL', ''DISGUST', 'SURPRISED'

720 SAMPLES (RESAMPLED)

CREMA-D

'NEUTRAL', 'CALM', 'HAPPY', 'SAD', "ANGRY', 'FEARFUL', ''DISGUST', 'SURPRISED'

15252 SAMPLES (RESAMPLED)



CREMA-D

CREMA-D is an emotional multimodal actor data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

Participants rated the emotion and emotion levels based on the combined audiovisual presentation, the video alone, and the audio alone. Due to the large number of ratings needed, this effort was crowd-sourced and a total of 2443 participants each rated 90 unique clips, 30 audio, 30 visual, and 30 audio-visual. 95% of the clips have more than 7 ratings.

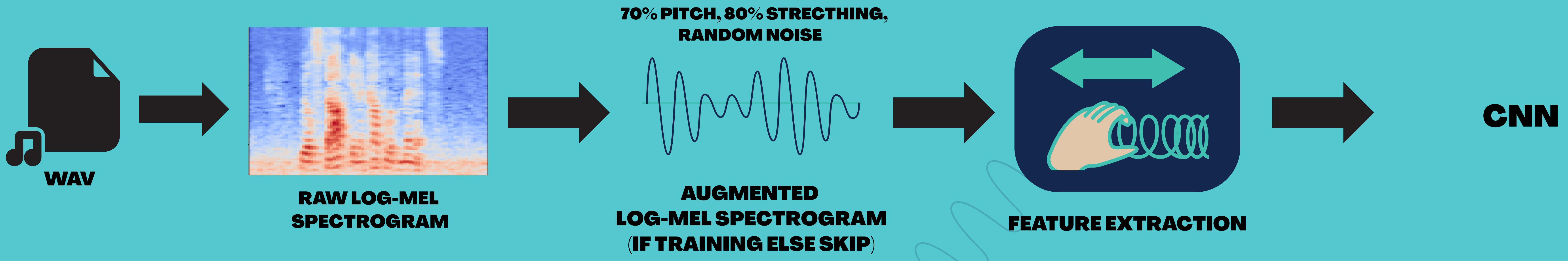
DATASETS

HAPPY, SURPRISED

SAD, ANGRY, FEARFUL, DISGUST, FEAR

NEUTRAL

CNN (PREP)



WAV

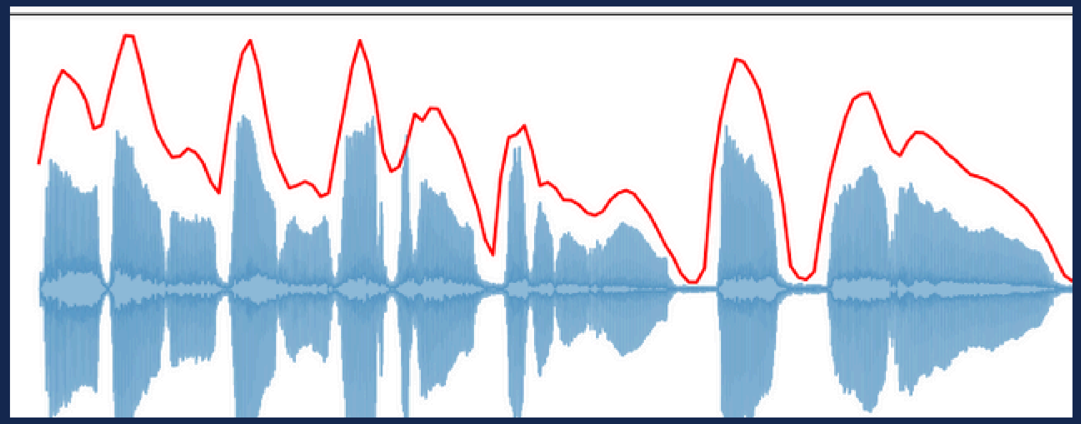
RAW LOG-MEL SPECTROGRAM

70% PITCH, 80% STRECTHING, RANDOM NOISE

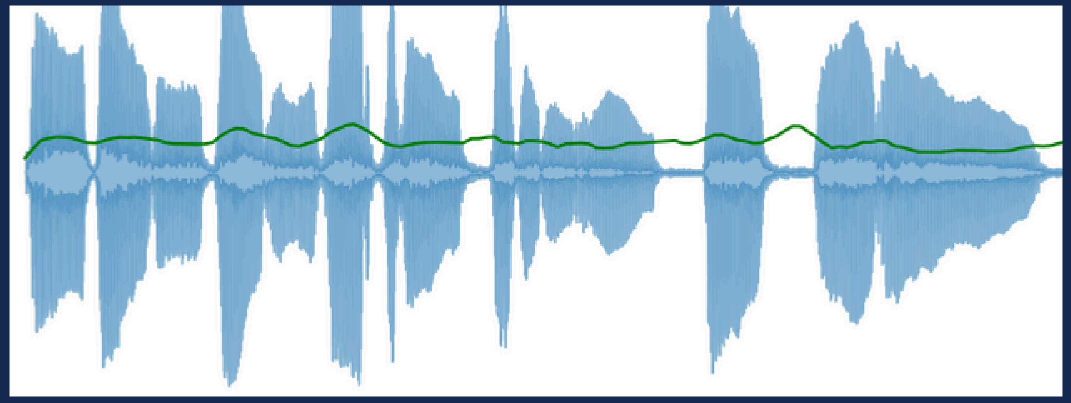
AUGMENTED LOG-MEL SPECTROGRAM (IF TRAINING ELSE SKIP)

FEATURE EXTRACTION

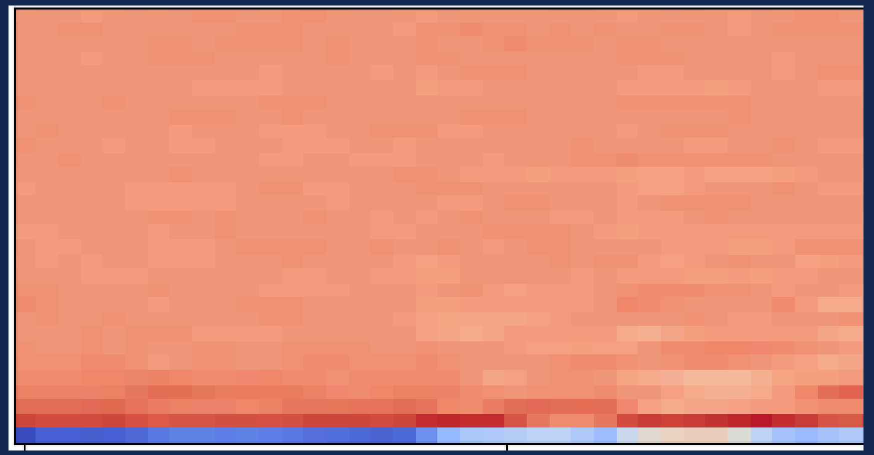
CNN



ZERO CROSSING RATE



ROOT MEAN SQUARED



MEL-FREQUENCY CEPSTRAL COEFFICIENTS

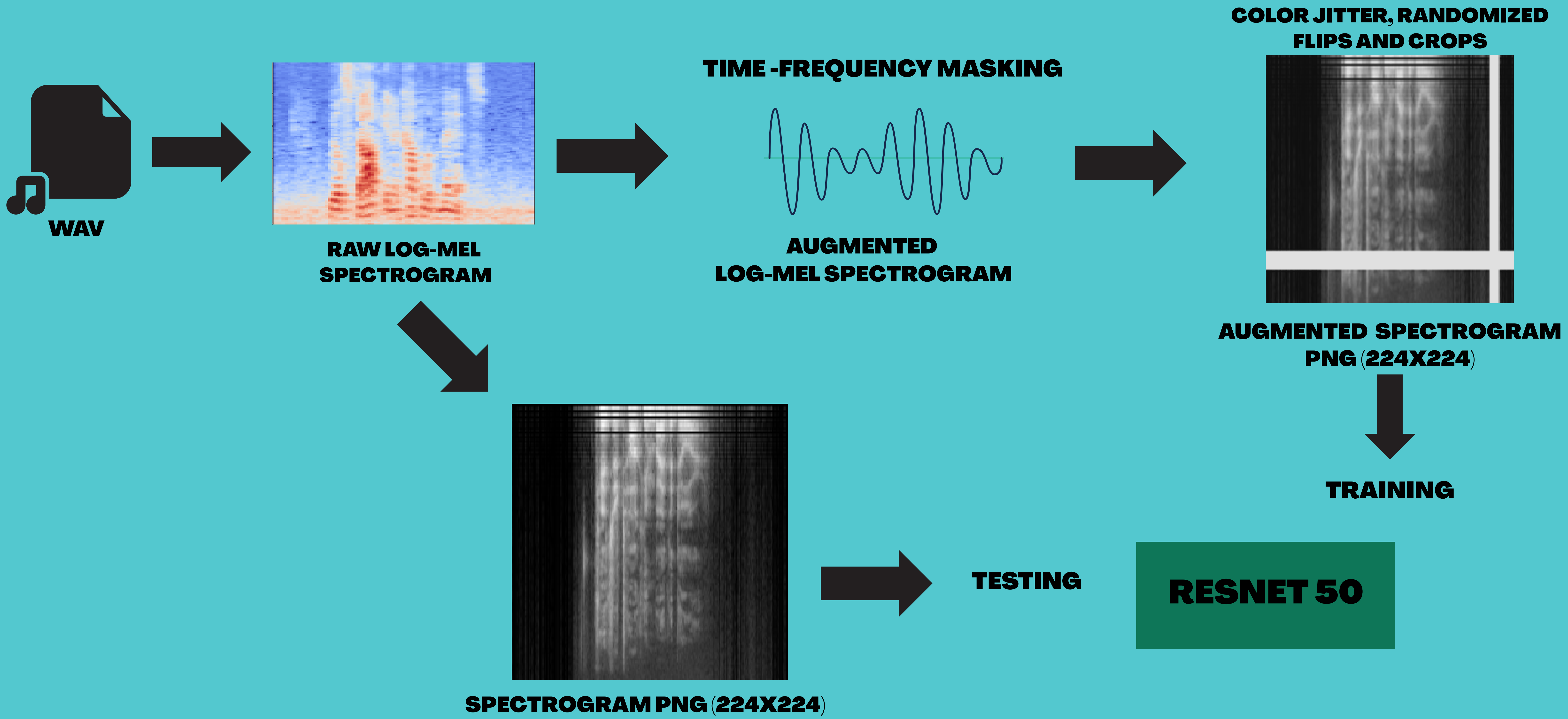
CNN

Layer Block	Description	Output Shape (Example)
Conv1 + BN + Pool (x2)	Conv1d layers with BatchNorm and MaxPool reduce time dimension by 4×	[B, 512, T/4]
Dropout	Regularization (p=0.2)	[B, 512, T/4]
Conv1 + BN + Pool (x3)	Conv1d layers with BatchNorm and MaxPool further reduce time dimension by 8×	[B, 128, T/32]
Dropout	Regularization (p=0.2)	[B, 128, T/32]

CNN

Layer Block	Description	Output Shape (Example)
Flatten + FC + BN	Flatten features and fully connected layer to 512 units	[B, 512]
FC (Output)	Final fully connected layer to 3 output logits	[B, 3]

RESNET 50 (PREP)



WAV

RAW LOG-MEL SPECTROGRAM

TIME -FREQUENCY MASKING

AUGMENTED LOG-MEL SPECTROGRAM

COLOR JITTER, RANDOMIZED FLIPS AND CROPS

AUGMENTED SPECTROGRAM PNG (224X224)

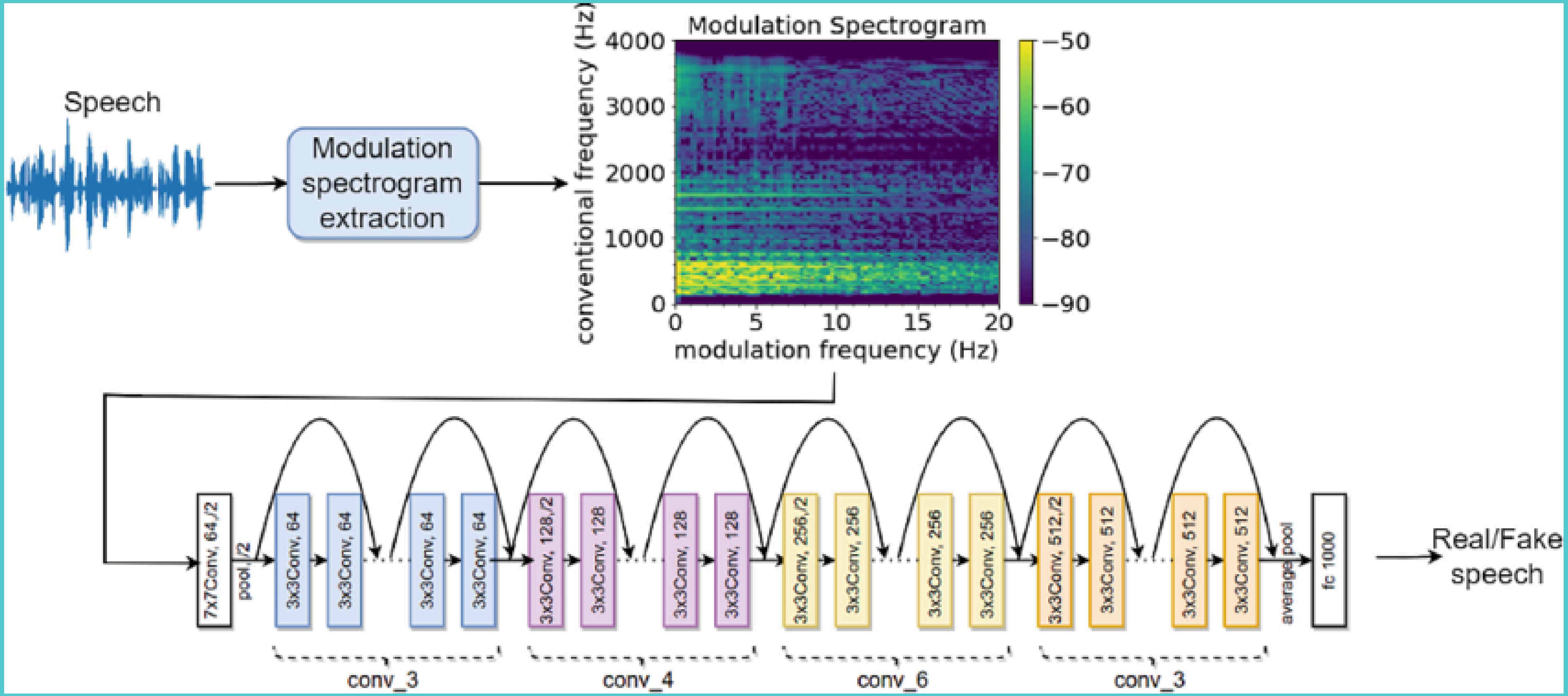
TRAINING

SPECTROGRAM PNG (224X224)

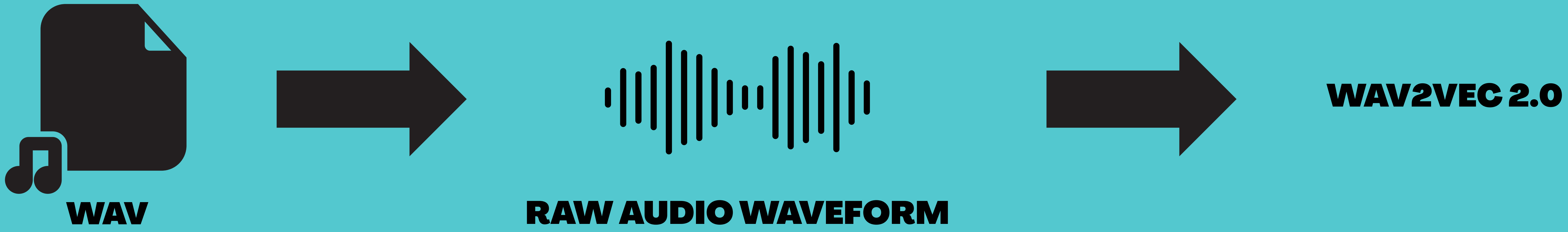
TESTING

RESNET 50

RESNET 50


Speech
Modulation spectrogram extraction
Modulation Spectrogram
conventional frequency (Hz)
4000
3000
2000
1000
0
modulation frequency (Hz)
0
5
10
15
20
−50
−60
−70
−80
−90
7x7Conv, 64,/2
pool,/2
3x3Conv, 64
3x3Conv, 64
3x3Conv, 64
3x3Conv, 64
3x3Conv, 128,/2
3x3Conv, 128
3x3Conv, 128
3x3Conv, 128
3x3Conv, 256,/2
3x3Conv, 256
3x3Conv, 256
3x3Conv, 256
3x3Conv, 512,/2
3x3Conv, 512
3x3Conv, 512
3x3Conv, 512
average pool
fc 1000
Real/Fake speech
conv_3
conv_4
conv_6
conv_3

WAV2VEC 2.0 (PREP)

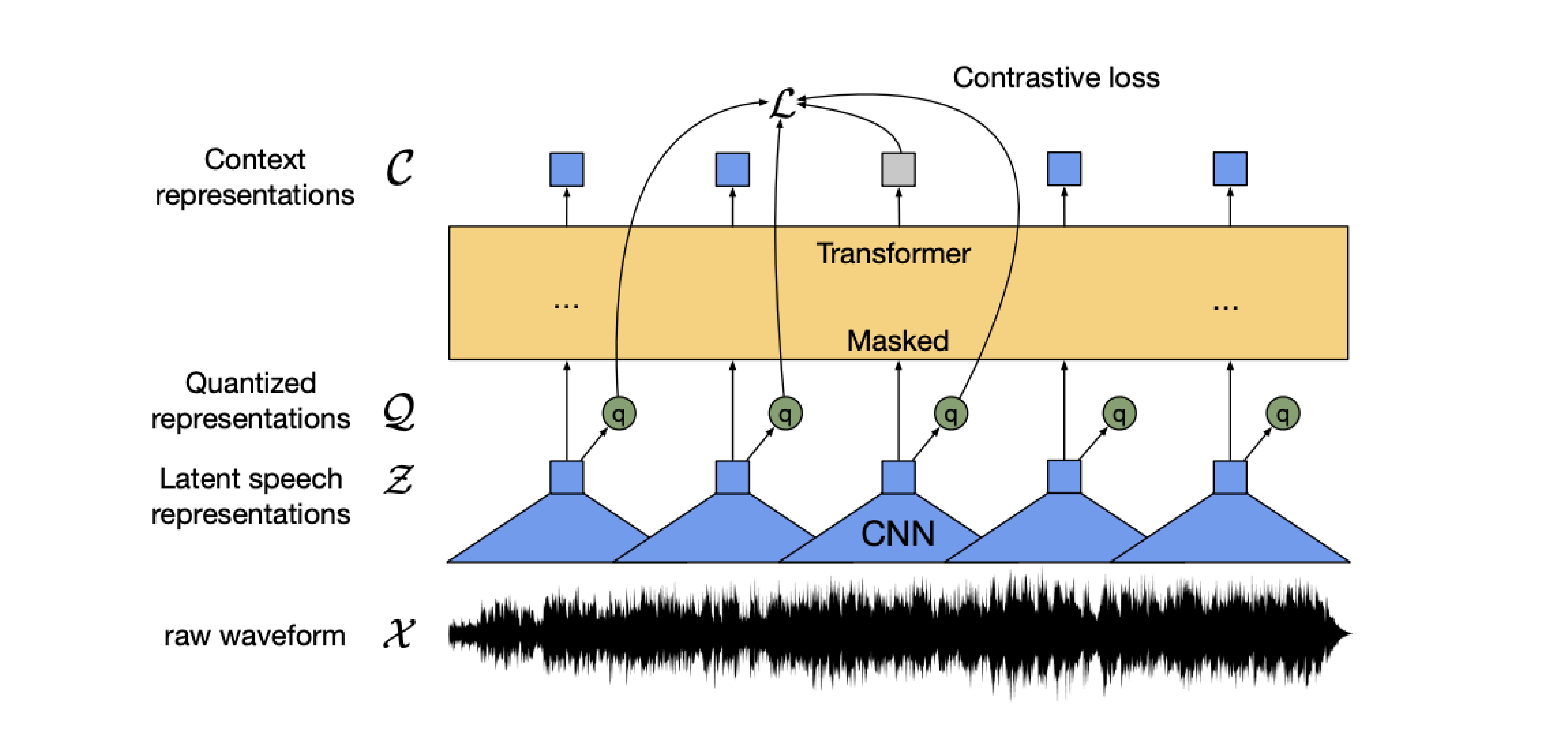


WAV

RAW AUDIO WAVEFORM

WAV2VEC 2.0

WAV2VEC 2.0


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
...
...
Masked
Quantized representations
$\mathcal{Q}$
q
q
q
q
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

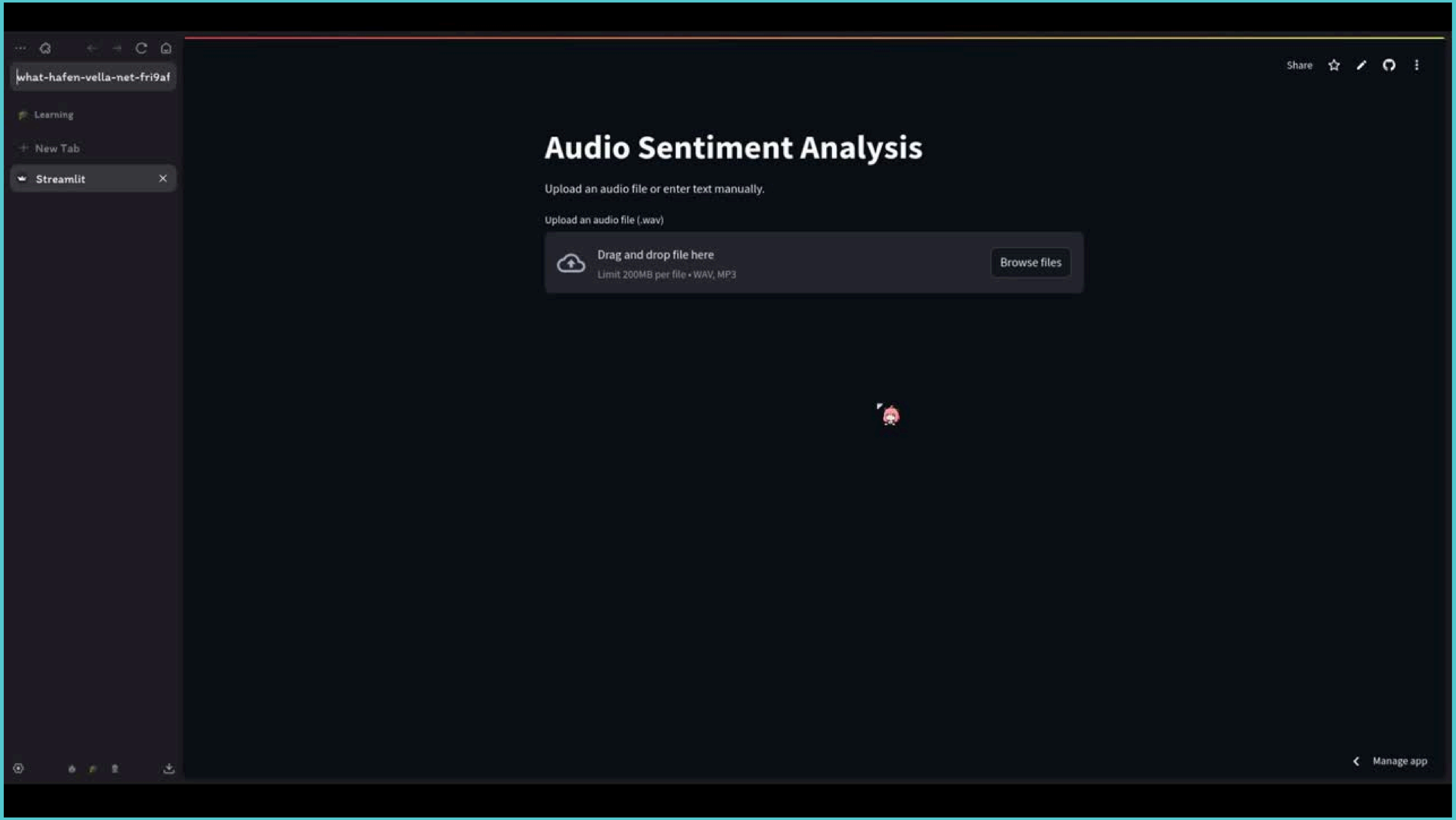
MODEL PERFORMANCE

Metrics	CNN	ResNet50	Wav2Vec 2.0
Epoch	25	10	5
Accuracy	0.9627	0.6562	0.9762
F1	0.9625	0.6604	0.9762
Cohen's Kappa	0.9441	0.4843	0.9642

CONCLUSION

- For the text-based models, RoBERTa performed the best overall, reaching 92% accuracy and a strong F1 score of 0.91.
- On the audio side, Wav2Vec 2.0 delivered the highest performance among all models tested. It achieved 97.6% accuracy and a Cohen’s Kappa of 0.964,
- both text and audio models have their strengths. Text-based models like RoBERTa are excellent when clean, accurate transcriptions are available. However, audio-based models, especially Wav2Vec 2.0, can recognize emotions directly from voice tone, making them more flexible in situations where the exact words are unclear or unimportant.

VIDEO PRESENTATION



Audio Sentiment Analysis

Upload an audio file or enter text manually.

Upload an audio file (.wav)

Drag and drop file here

Limit 200MB per file • WAV, MP3

Browse files

SUSTAINABLE DEVELOPMENT GOAL

Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.

8.5.1

AVERAGE HOURLY EARNINGS OF FEMALE AND MALE EMPLOYEES, BY OCCUPATION, AGE AND PERSONS WITH DISABILITIES

8.5.2

UNEMPLOYMENT RATE, BY SEX, AGE AND PERSONS WITH DISABILITIES

8.5.1

THANK YOU FOR LISTENING !

TEAM 13

Dejoras, Dylan James
Villamor, Kurt Russel A.