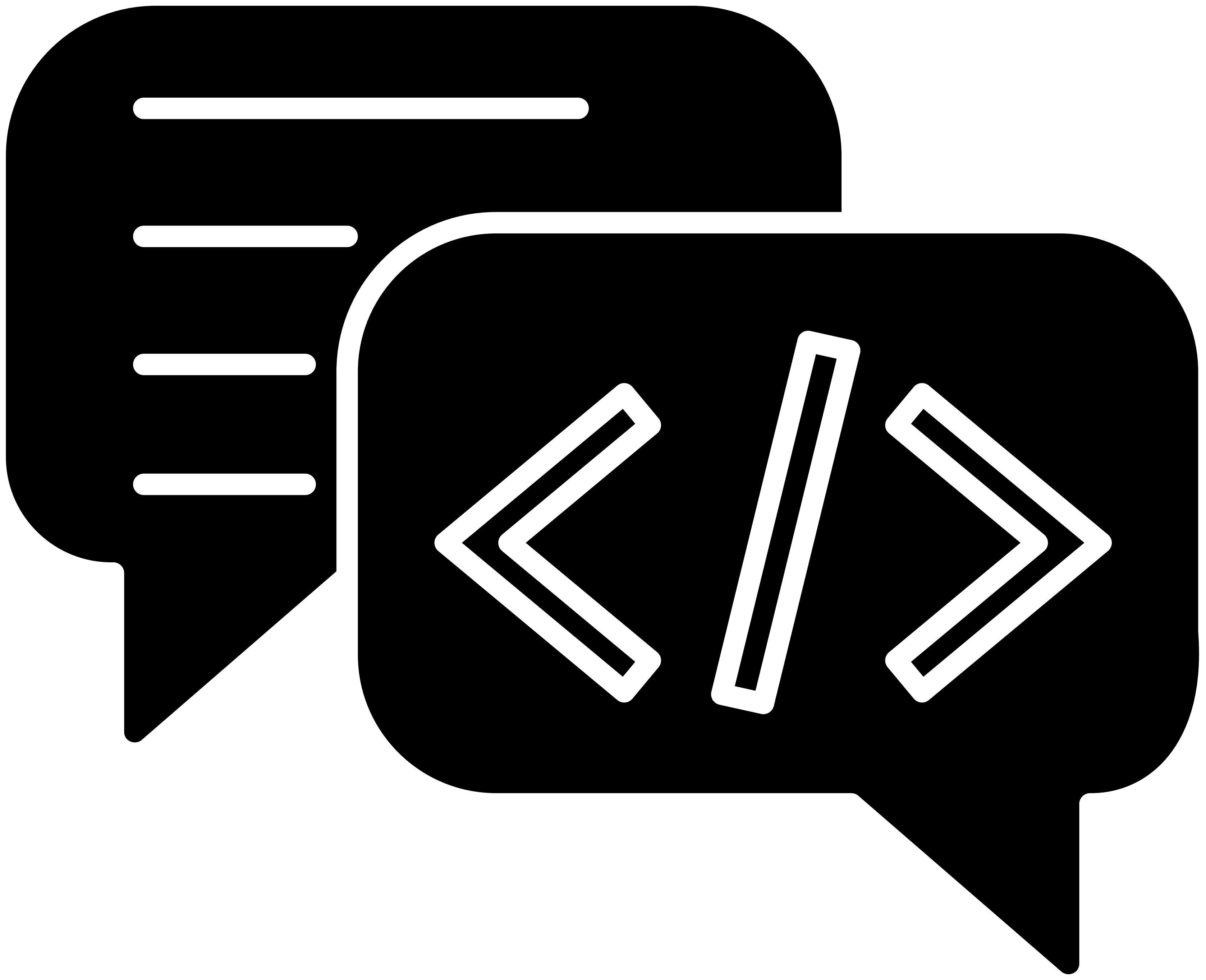
Univ.AI

Amigo

DialoGPT-based Chatbot that listens to you and your emotion.

AI3 Project By:
Anshika, Niegil, Sakthisree, Vishnu



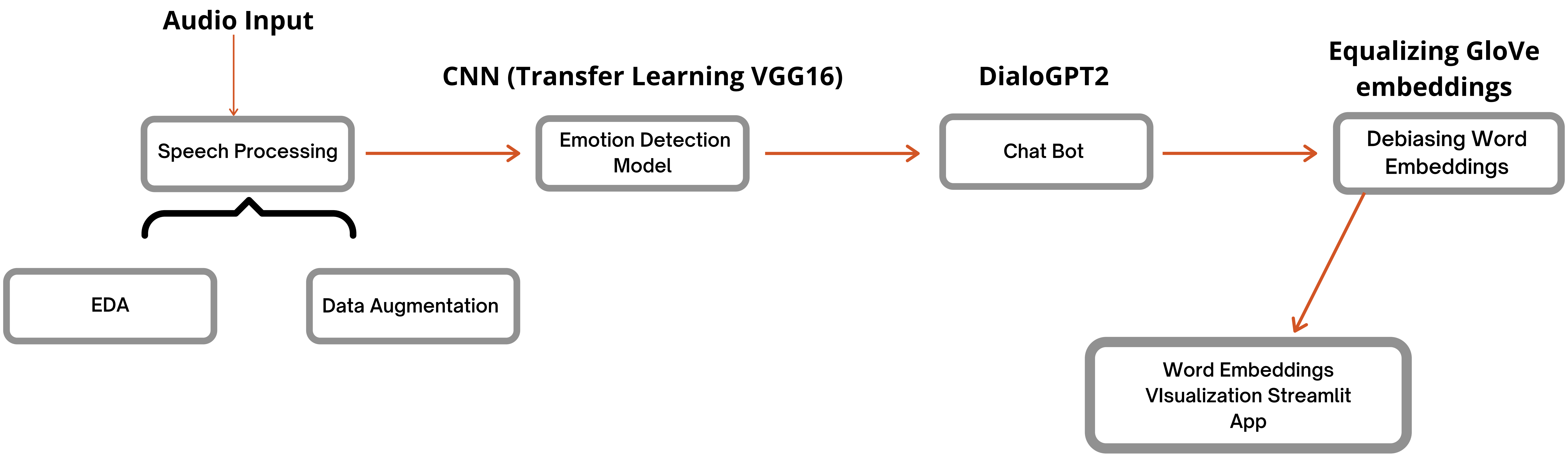
Introduction

Describe The Problem

1. Most chat-bots that we have seen so far extract emotion from the text. Missing out on the tone in which the statement is put across. To tackle this, we have developed a chatbot that responds to the most important aspect of communication, ie. the tone.
2. Another challenge is bias in text data. In order to curb so, we looked at debiasing word embeddings from GLoVe to understand methodologies.

Project Flow

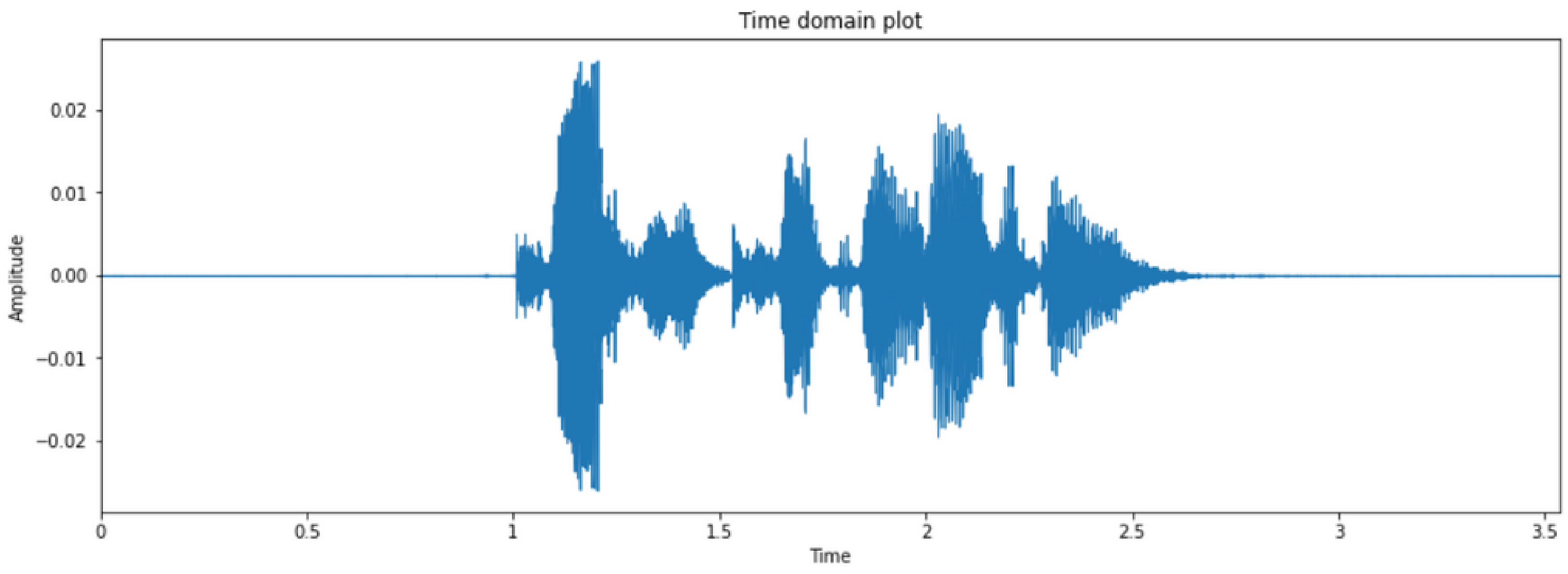
How our project was structured and the various models used?



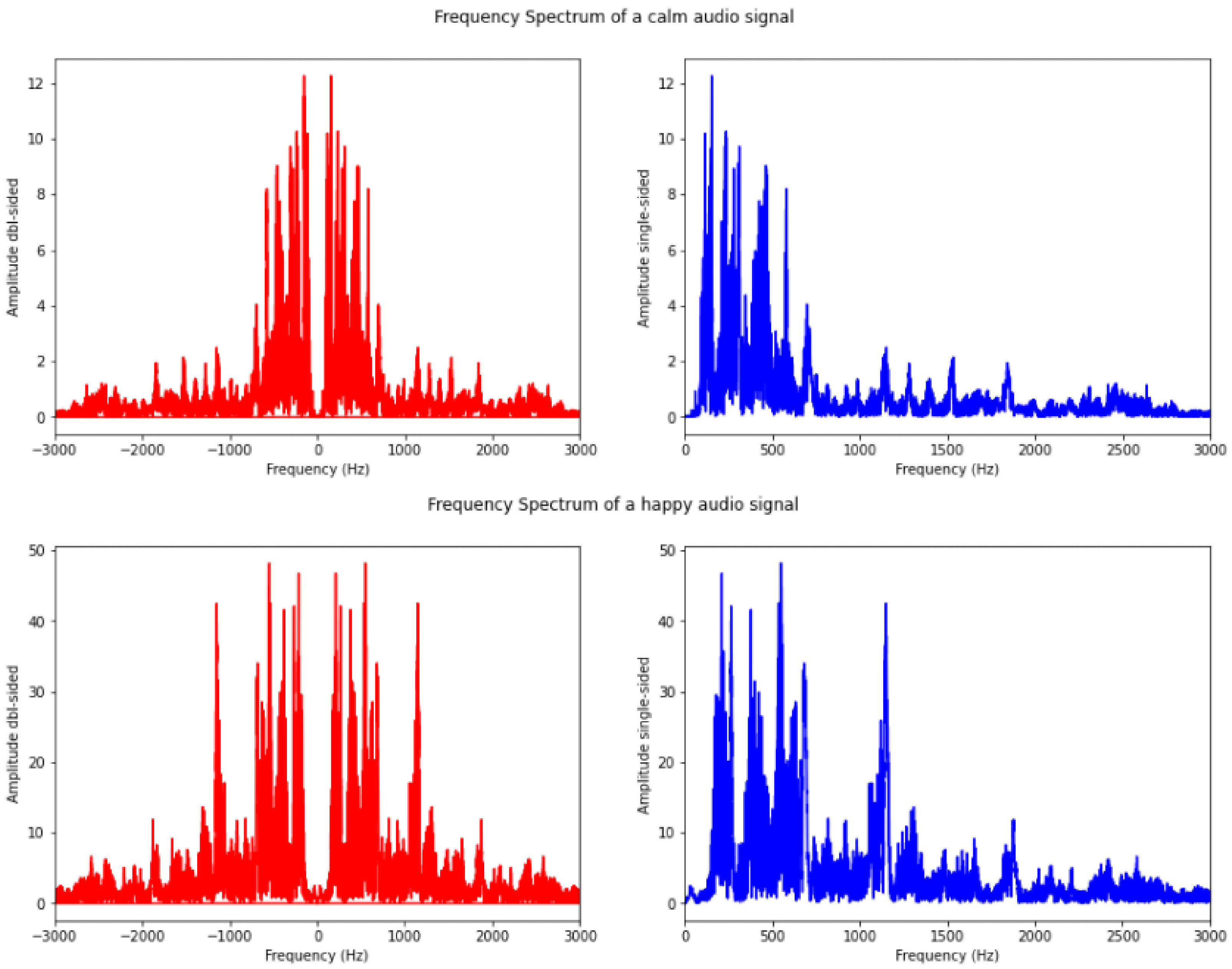
Speech Processing

Exploratory Data Analysis

Time Domain



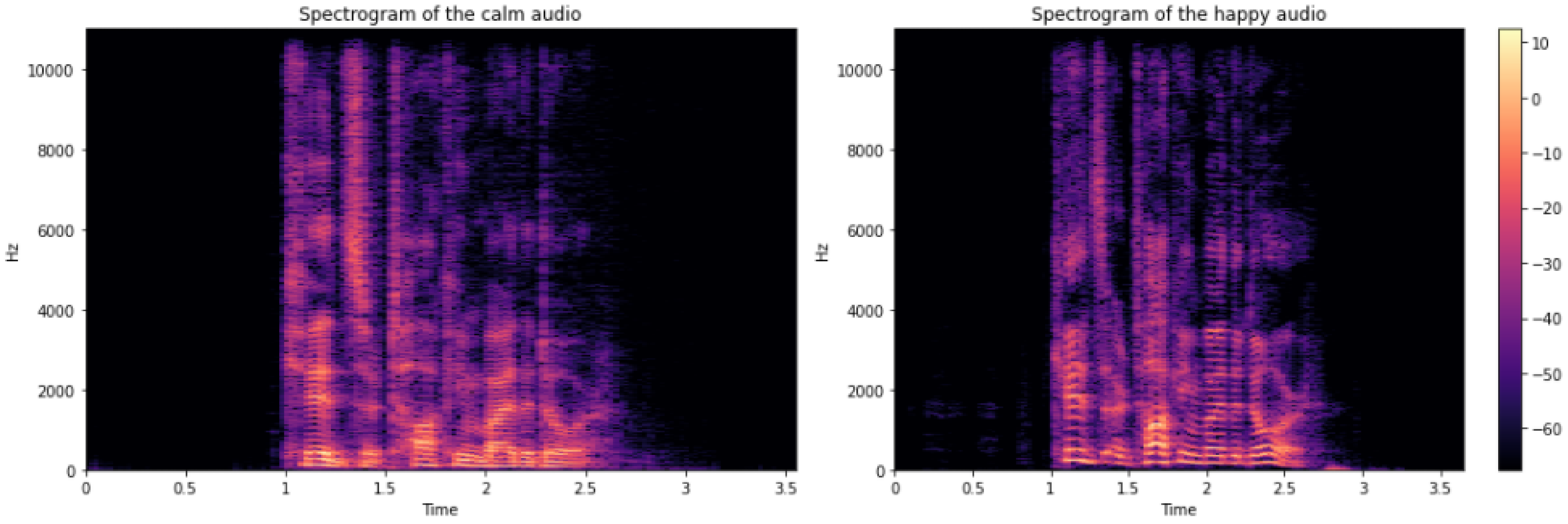
Frequency Domain



Speech Processing

Exploratory Data Analysis

Time-Frequency Domain



Speech Processing

Data Augmentation

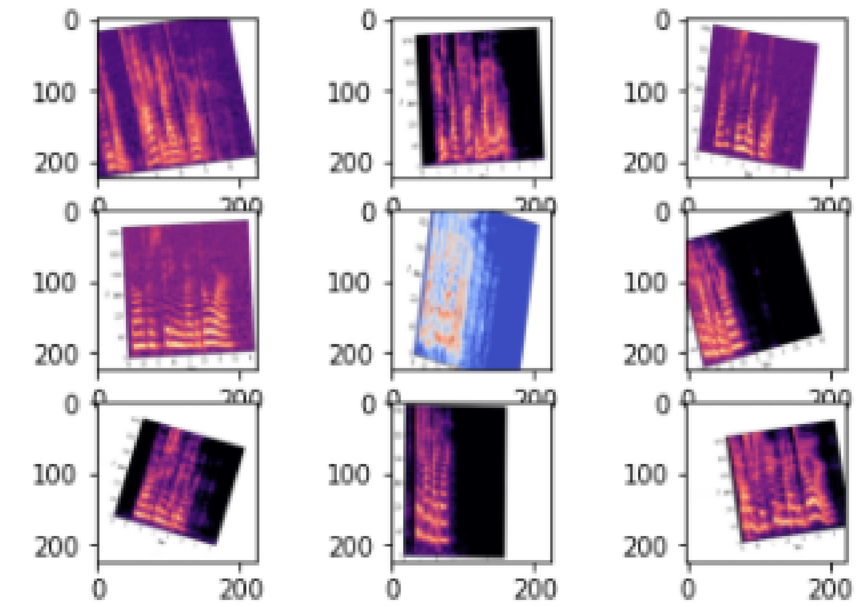
Signal Augmentation

Original Signal
Stretched Signal
White Noise added Signal
Compressed Signal

Spectrogram →

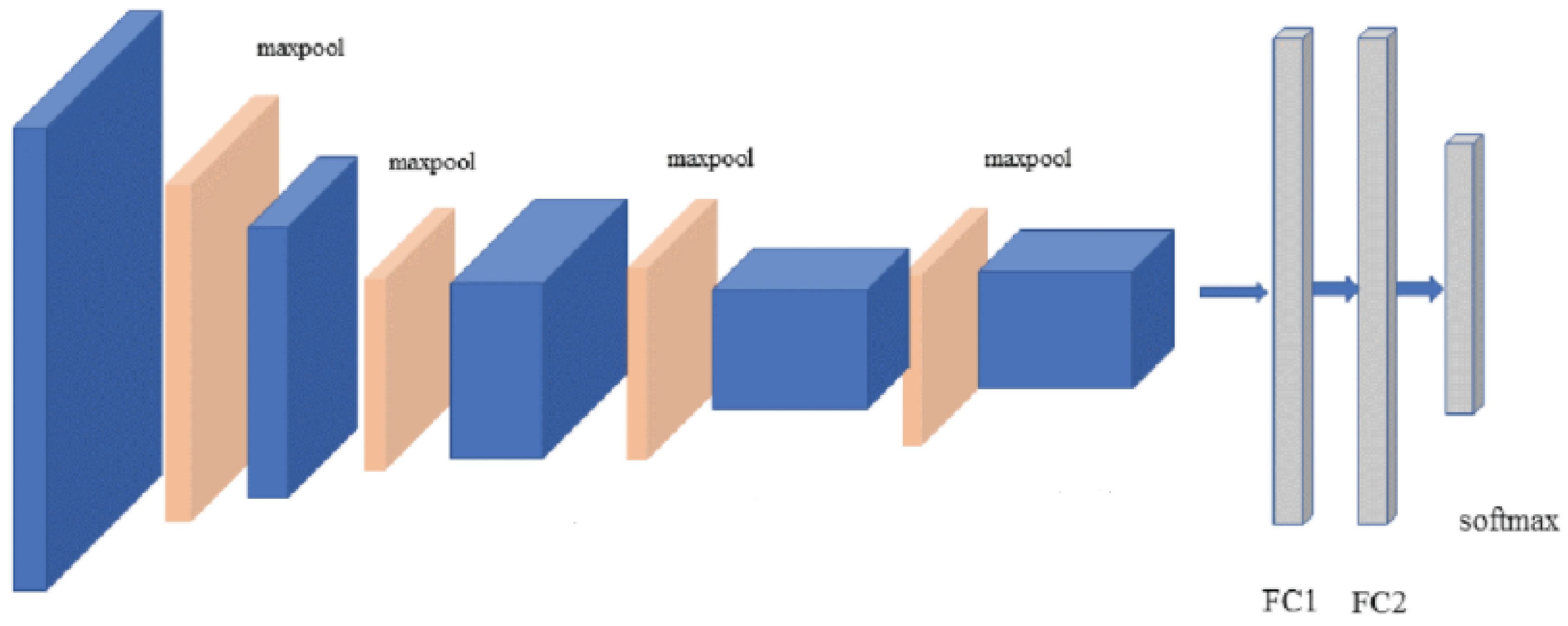
Image Augmentation

Rotation
Zoom
Width Shift



Emotion Detector using Transfer Learning

VGG16 Architecture



Dataset used for Training

The Ryerson Audio-Visual Database of Emotional...

Citing the RAVDESS The RAVDESS ...

zenodo.org

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition”. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes.

Model: "model"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 512)	12845568
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 256)	131328
dense_2 (Dense)	(None, 8)	2056

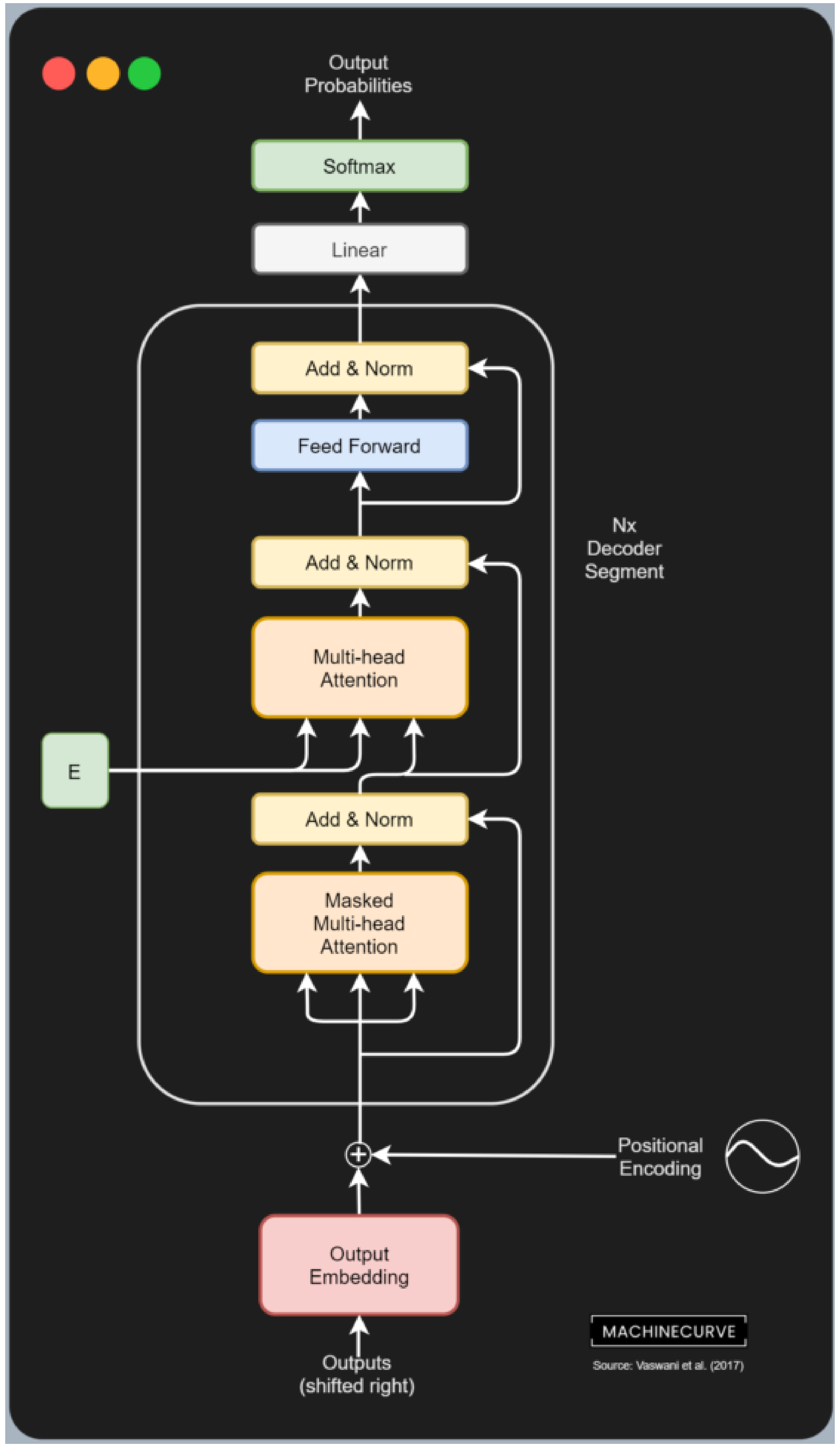
Total params: 27,693,640
Trainable params: 20,058,376
Non-trainable params: 7,635,264

DialoGPT-2

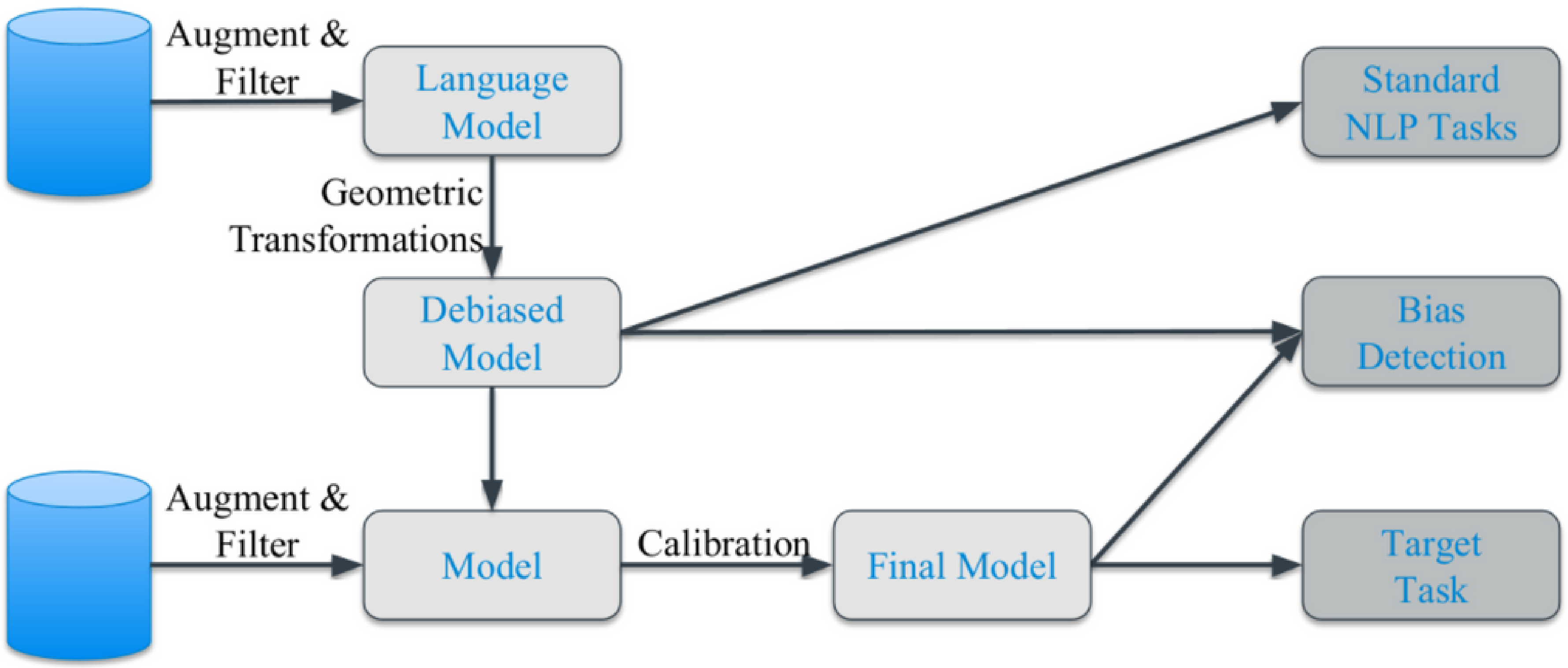
DialoGPT was trained with a causal language modeling (CLM) objective on conversational data and is therefore powerful at response generation in open-domain dialogue systems.
rained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017.

We leveraged DialoGPT generate more relevant, contentful and context-consistent responses than strong baseline systems.

The pre-trained model and training pipeline are publicly released to facilitate research into neural response generation and the development of more intelligent open-domain dialogue systems.



Debiasing the Word Embeddings



$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to

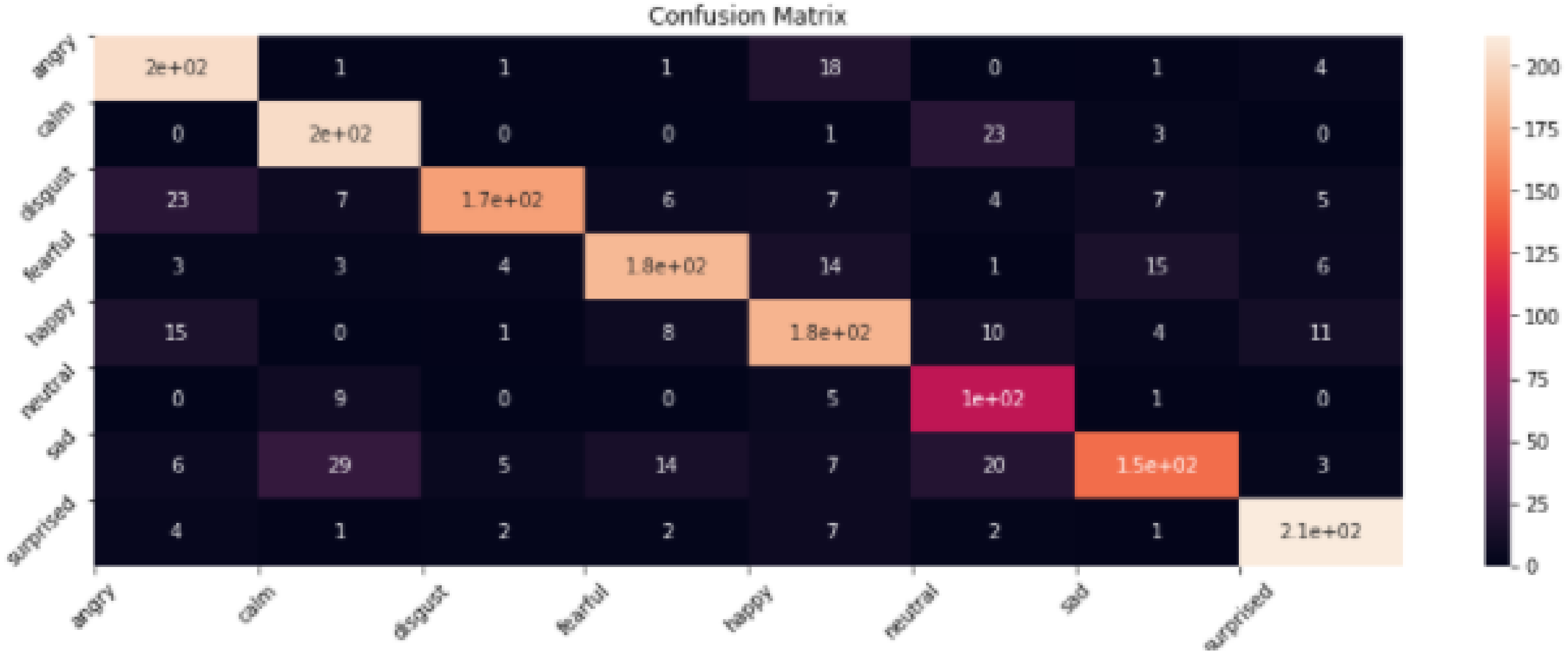
$$\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\|.$$

For each set $E \in \mathcal{E}$, let

$$\begin{aligned} \mu &:= \sum_{w \in E} w/|E| \\ \nu &:= \mu - \mu_B \\ \text{For each } w \in E,\quad \vec{w} &:= \nu + \sqrt{1 - \|\nu\|^2}\,\frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|} \end{aligned}$$

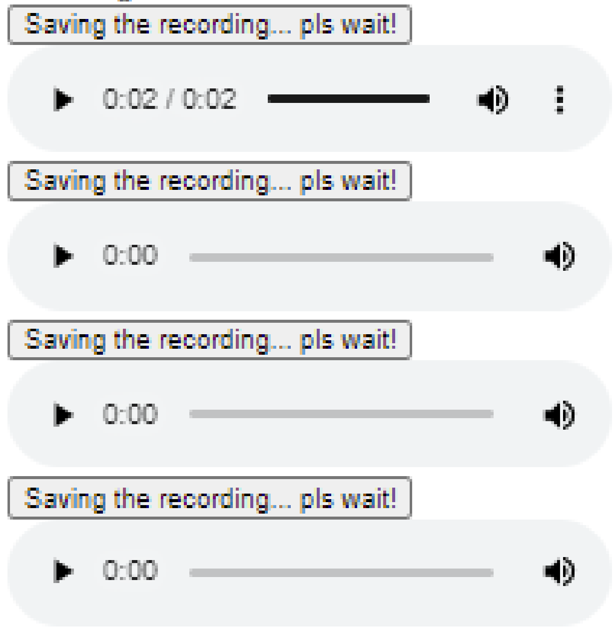
The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases

Confusion Matrix of CNN detecting Emotion



	precision	recall	f1-score	support
angry	0.80	0.89	0.84	230
calm	0.80	0.88	0.84	230
disgust	0.93	0.74	0.83	230
fearful	0.86	0.80	0.83	230
happy	0.76	0.79	0.77	231
neutral	0.62	0.87	0.73	115
sad	0.82	0.64	0.72	231
surprised	0.88	0.92	0.90	231
accuracy			0.81	1728
macro avg	0.81	0.82	0.81	1728
weighted avg	0.82	0.81	0.81	1728

```
>> You: how are you
Emotion detected  happy
>> DialoGPT: Hey hey, someone's in a good mood.
I'm good, you?
>> You: how is the weather today I'm good
Emotion detected  fearful
>> DialoGPT: Don't worry! I am not Donald Trump.
It's nice, you?
>> You: that's good to hear
Emotion detected  angry
>> DialoGPT: Having a bad day? You seem to be angry
that's good to hear
>> You: I'm not really that angry
Emotion detected  happy
>> DialoGPT: Hey hey, someone's in a good mood.
that's good to hear
```




```
>> You: what did the doctor tell the receptionist
Emotion detected  angry
DialoGPT: Having a bad day? You seem to be angry
I don't know, but I think it was something about the doctor's wife.
```

Our Results

Chatbot (emotion CNN +DialoGPT2)

Before Debiasing

```
cosine similarities before equalizing:
cosine_similarity(word_to_vec_map["man"], gender) =  -0.11711095765336832
cosine_similarity(word_to_vec_map["woman"], gender) =  0.35666618846270376
```

Cosine Similarity to woman-man

Technology	-0.13193732447554302
Engineer	-0.0803928049452407
Doctor	0.11895289410935041
Grandfather	0.023629798450867857
Grandmother	0.3846014363741861
Literature	0.06472504433459932

After debiasing

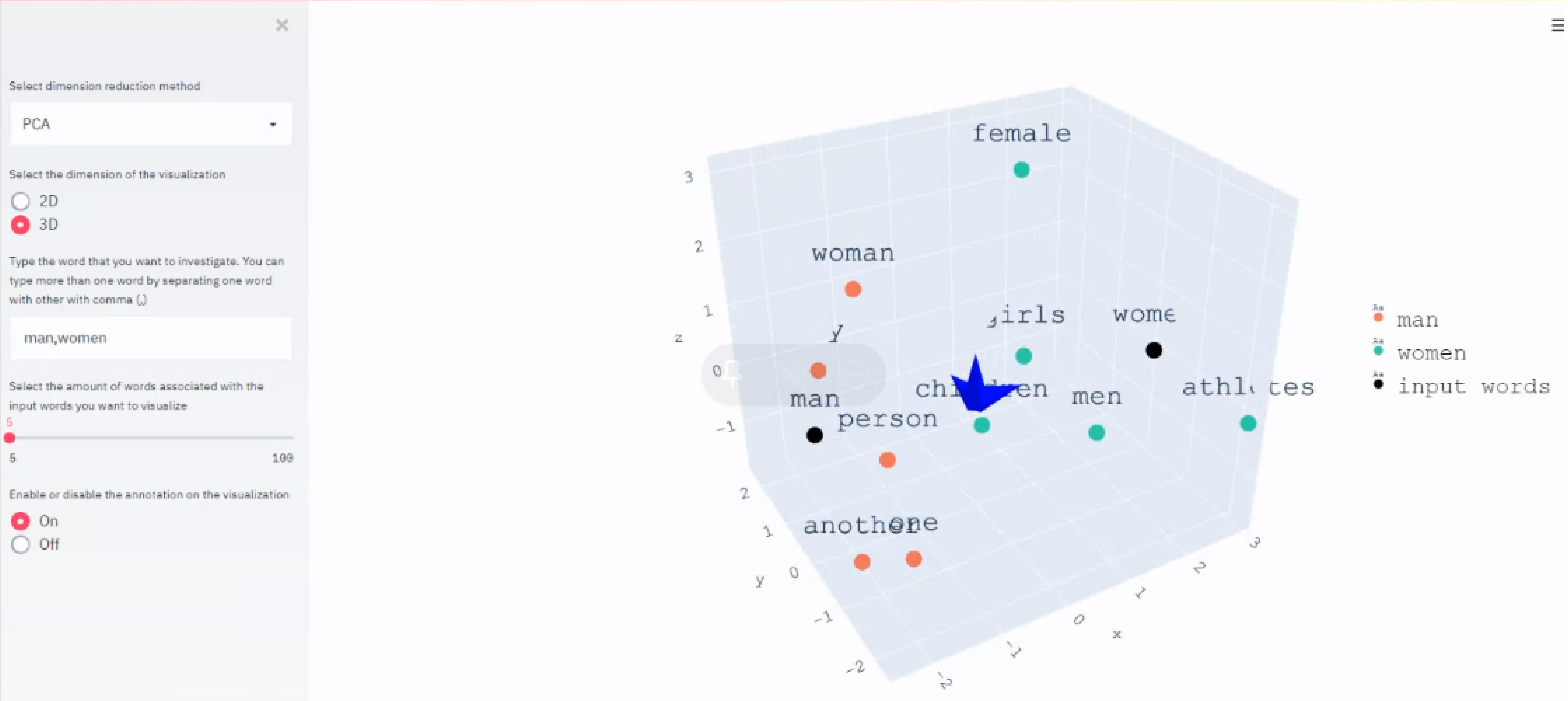
```
cosine similarities after equalizing:
cosine_similarity(e1, gender) =  -0.7165727525843937
cosine_similarity(e2, gender) =  0.739659647492891
```

Cosine Similarity to woman-man

Technology	0.043615821441082496
Engineer	-0.0064291345956580285
Doctor	-0.059556124904608376
Grandfather	-0.014652493379103636
Grandmother	-0.24509001213297088
Literature	-0.08286463249952107

Our Results

Debiasing Results



Our Results

Streamlit App

Further Work

- Debasing contextual and positional embeddings outside of GLoVe.
- Developing interface with chatbot.
- Fine tuning DialoGPT further
- Application of the same in the area of Mental Health - to assess tone of person using chatbot and accordingly provide audio responses catered to the mental health issue.