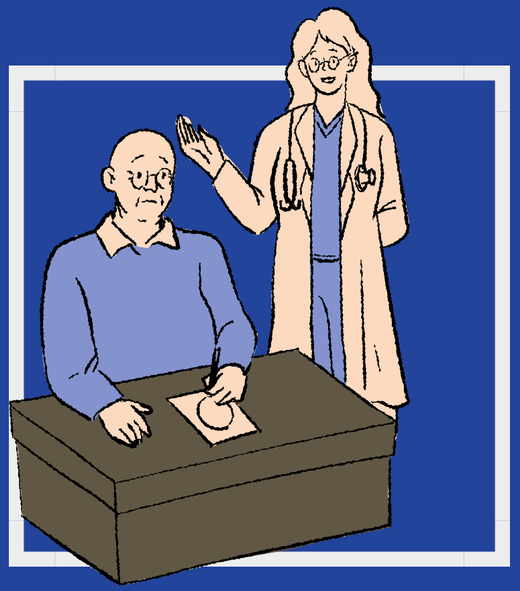
Dementia Classification based on Linguistic Analysis

Yunxi (Claire) Cheng, Tripti Sharma, Linting (Linsey) Wang, Sifan(Ellen) Zhu

Agenda

1. Dataset Introduction
2. LSTM without Embedding
3. LSTM with Embedding
4. Attention Model
5. Model Comparison
6. Conclusion and Future Research



Dementia affects over 50 million people worldwide, and this number is expected to triple by 2050.

Dementia affects around 10% of people over the age of 65 and approximately 50% of those over 85 years old.

The data in DementiaBank has been annotated with information about the speakers, the tasks performed, and the linguistic features of the data.

About the Dataset

Data source description

- **DementiaBank**: a shared database for the study of communication in dementia.
- Participants are **recorded** describing given topics.
- **Scripts** are transcribed by computer with the CLAN program into CHAT format.

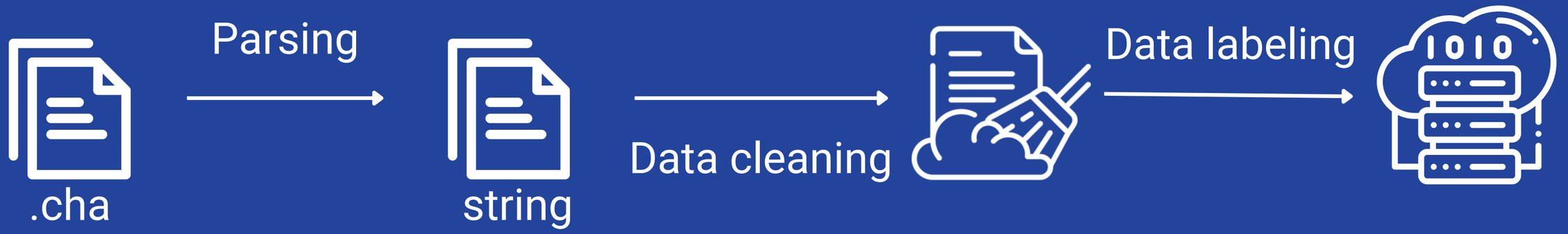
https://dementia.talkbank.org/access/English/Pitt.html

Letters	Categories	Example	Meaning	POS
@a	addition	xxx@a	unintelligible	w
@b	babbling	abame@b	-	bab
@c	child-invented form	gumma@c	sticky	chi
@d	dialect form	younz@d	you	dia
@e	echolalia, repetition	want@e more@e	want more	skip
@f	family-specific form	bunko@f	broken	fam
@fp	filled pause	um@fp	use &-um instead	skip
@g	general special form	gongga@g	-	skip
@i	interjection, interaction	uhhuh@i	-	co
@k	multiple letters	ka@k	Japanese “ka”	n:let
@l	letter	b@l	letter b	n:let
@lp	letter plural	a@lp	several “a”s	n:let
@n	neologism	breaked@n	broke	neo
@o	onomatopoeia	woofwoof@o	dog barking	on
@p	phonol. consistent form	aga@p	-	phon
@q	metalinguistic use	no if@q-s or but@q-s	when citing words	meta
@s:*	second-language form	istenem@s:hu	Hungarian word	L2
@s$n	second-language noun	perro@s$n	Spanish noun	n
@si	singing	lalala@si	singing	sing
@sl	signed language	apple@sl	apple	sign
@sas	sign & speech	apple@sas	apple and sign	sas
@t	test word	wug@t	test	test
@u	Unibet transcription	binga@u	-	uni
@wp	word play	goobarumba@wp	-	wp
@x:*	excluded words	stuff@x	-	unk
@z:xxx	user-defined code	word@z:rtfd	any user code	

Column description

1) **speech** - The transcript
2) **has_dementia** - If the patient has dementia or not (0 - , 1-)

Data Cleaning and Preprocessing

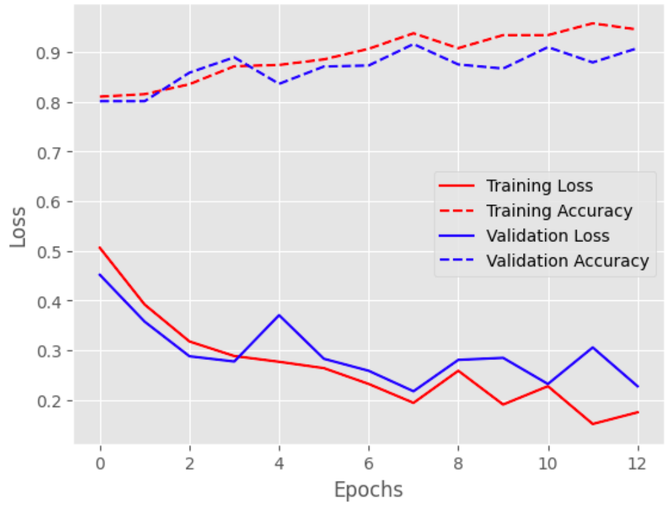


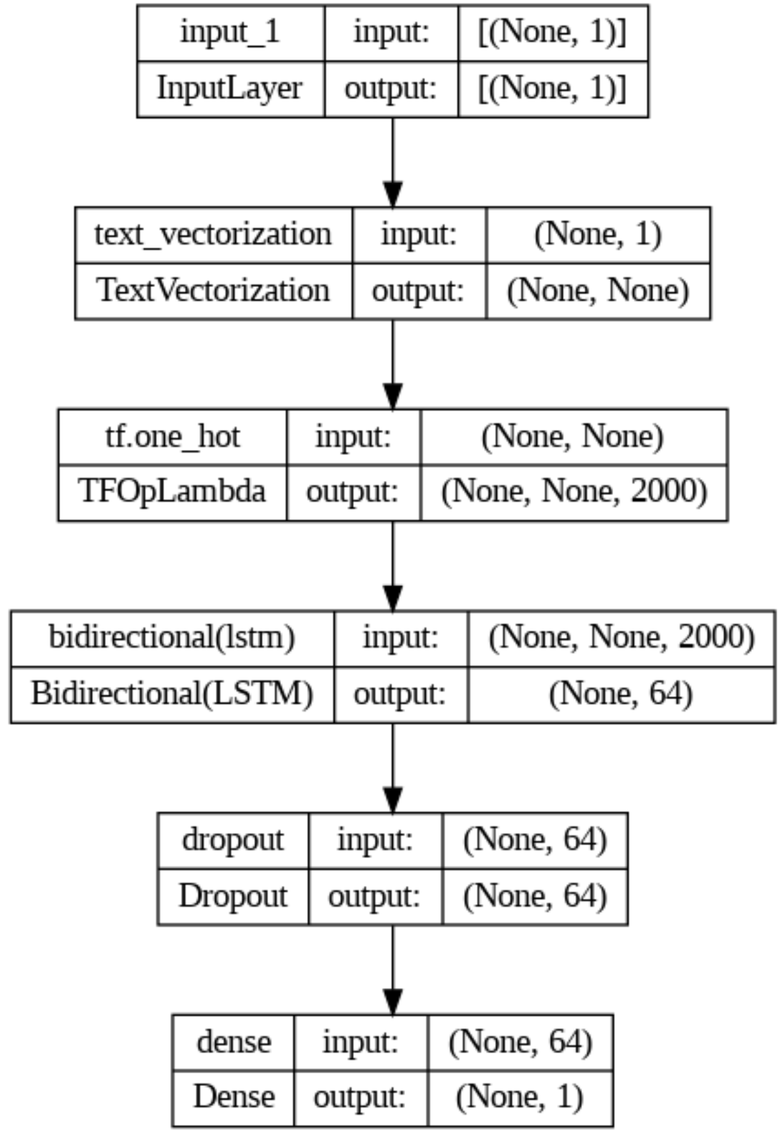
	speech	has_dementia
0	the boy is on a stool that is falling while he...	0
1	+< oh . [+ exc] +< cat [/] cat . [+ exc] ...	1
2	I write with a pencil . that was what ? [+ ...	1
3	this little girl <went to the> [//] &-uh went ...	1
4	hm . [+ exc] xxx . [+ exc] this little boy ...	0

total samples: 1287
total control: 247
total dementia: 1040

LSTM Model

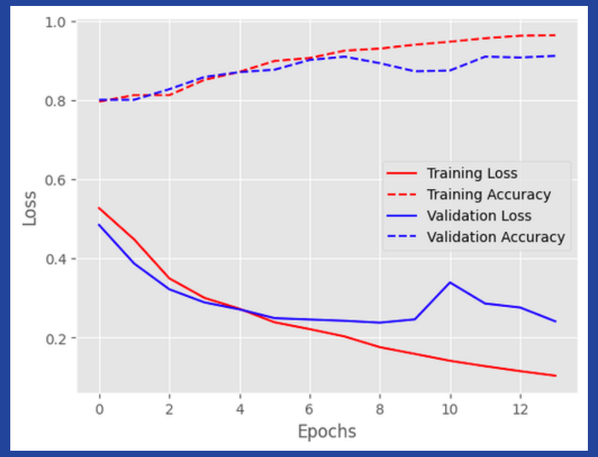
- Text_vectorization (top 2000 tokens)
- One_hot encoding layer
- LSTM layer with 32 LSTMs
- EarlyStopping with patience=5
- Dropout with 0.5 dropout rate
- Output with sigmoid activation function
- Binary cross-entropy loss

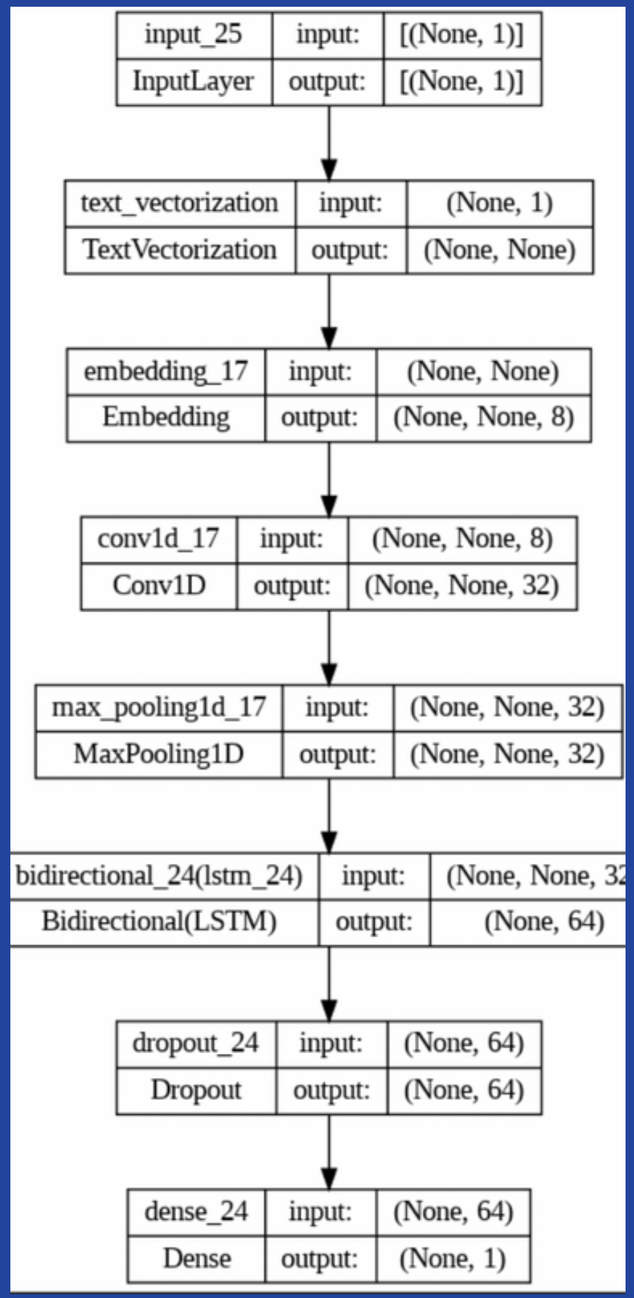




LSTM with 1D Convolution Layer for Embedding

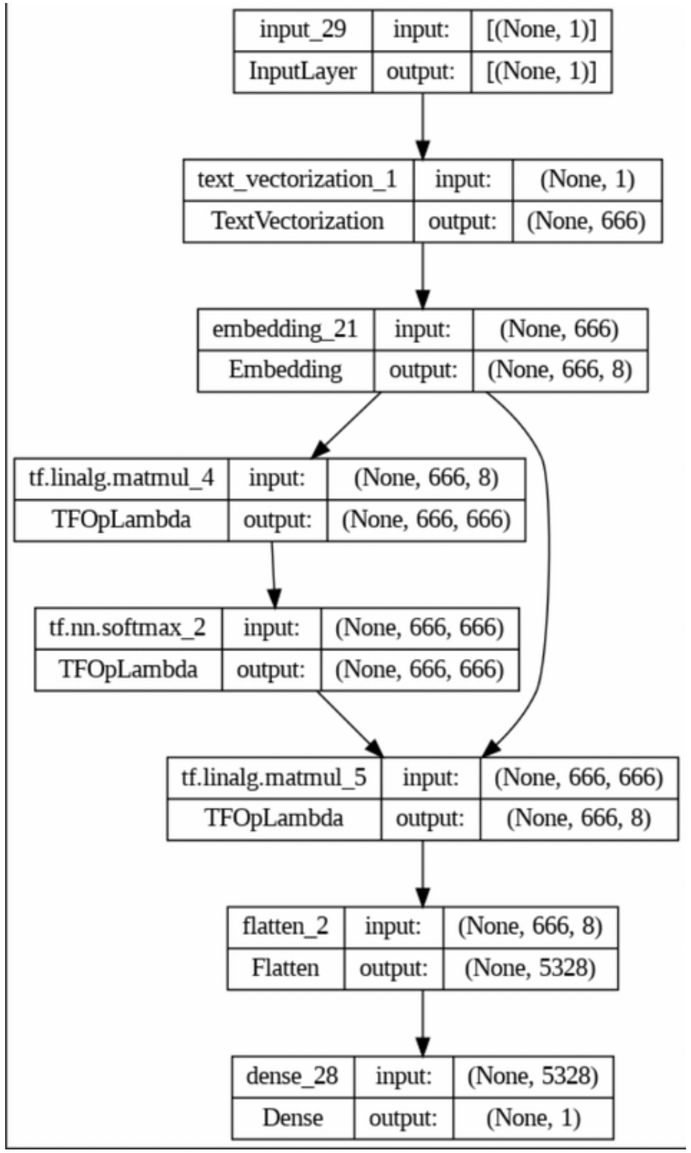
- Text_vectorization layer (top 2000 tokens)
- One-hot encoding layer
- An embedding layer with a dimension of 8
- A 1D convolutional layer and pooling layer
- A bidirectional LSTM layer
- Output with sigmoid activation function
- Compiled with the RMSprop optimizer and binary cross-entropy loss
- More false positive predictions



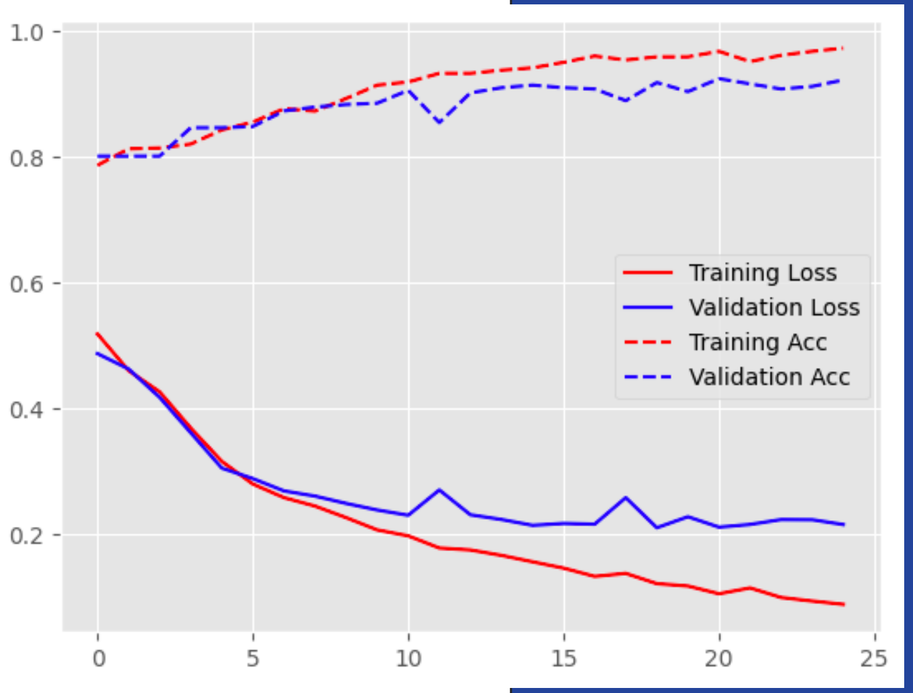


Self-Attention Model

- Text_vectorization layer
- Embedding layer with a dimension of 8
- A pairwise dot-product
- Softmax transformation to get the attention weights
- Sum and flatten weights
- Sigmoid activation function
- Adam optimizer and binary cross-entropy loss



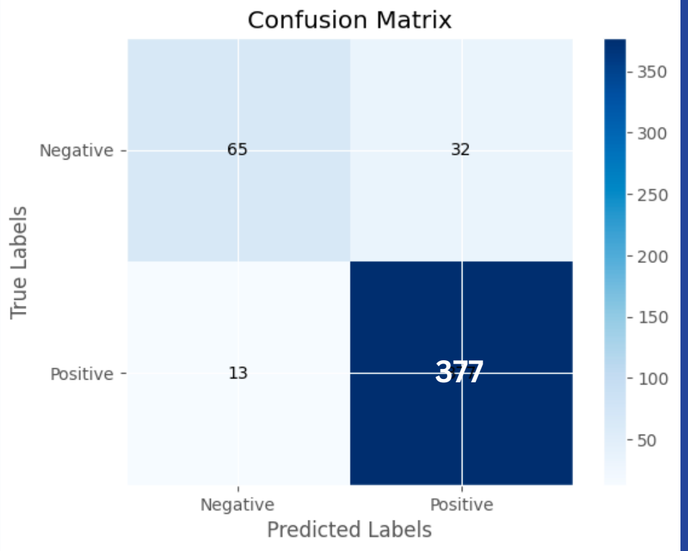
epochs=25, batch_size=25



Results Comparison

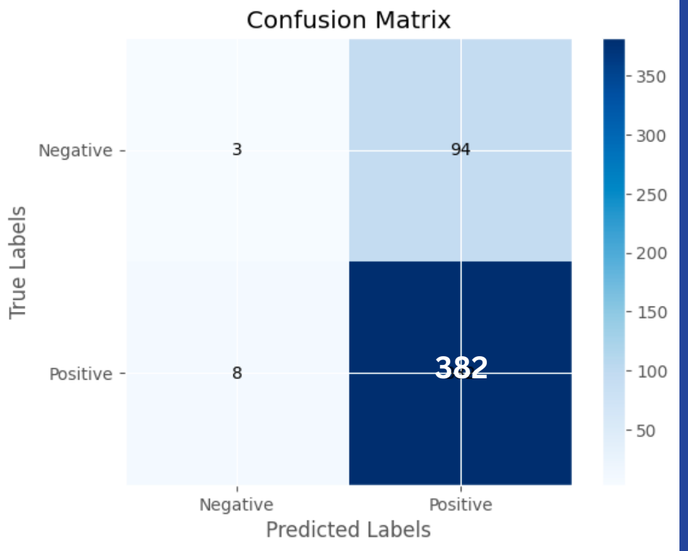
LSTM Model

- loss: 0.1751
- accuracy: 0.9450
- val_loss: 0.2275
- val_accuracy: 0.9076



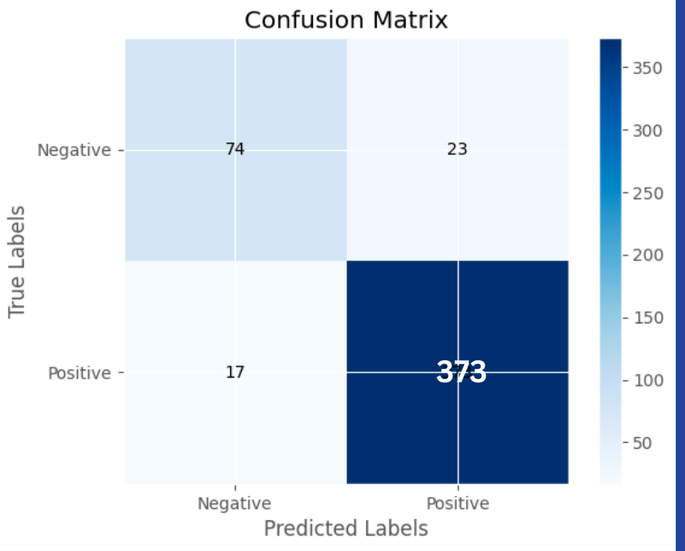
LSTM with Embedding

- loss: 0.1038
- accuracy: 0.9638
- val_loss: 0.2412
- val_accuracy: 0.9117



Attention Model

- loss: 0.0879
- accuracy: 0.9725
- val_loss: 0.2149
- val_accuracy: 0.9220



Conclusion and Future Work

- **Attention model** gives the most accurate classification for dementia.
- Our model can help healthcare provider to **diagnose** patients with dementia and allocate resources more efficiently.

- Future works can include:
  - Use **audio data** to avoid the required training for CLAN softwares.
  - Include more training data of **speech from daily dialogs**.
  - Leverage the developed model to classify different stages of dementia.

Thank you!