Multimodal Sarcasm Detection

Yufei Wang, Mona Gandhi, Zhen Huang, Haoxin Chen

Penn Engineering

Definition & Problem

Sarcasm

Oh, my God. What happened?
Oh, hum, Joey was born and then 28 years later,
I was robbed!

Sarcasm

Sarcasm is Difficult to Detect

- Delivery of positive sentiment in a negative scenario.
- It is essential for real world NLP applications like Chat-bot or AI customer service.
- Especially difficult from textual evidence alone.
- Detection can be made easier with the help of visual and audio cues.

Problem at Hand

Can we create a model that detects sarcasm by incorporating not only text data, but also audio and visual data?

Background

MUStARD (Castro et. al, 2019)

- A dataset specifically created for sarcasm detection
- Features a collection of YouTube videos from popular TV shows
- Total of 690 entries, with 345 labeled as sarcasm and 345 labeled as not sarcasm.

MUStARD (Castro et. al, 2019)


```
{
  "1_60": {
    "utterance": "It's just a privilege to watch your mind at work.",
    "speaker": "SHELDON",
    "context": [
      "I never would have identified the fingerprints of string theory in the aftermath of th
      "My apologies. What's your plan?"
    ],
    "context_speakers": [
      "LEONARD",
      "SHELDON"
    ],
    "sarcasm": true
  }
}
```

MUStARD (Castro et. al, 2019)

Text Features: BERT Embeddings

Audio Features: Speech processing library Librosa

Video Features: ResNet-152

SVM Model incorporating T+A+V:

Precision: 64.3%

Recall: 62.6%

F1-Score: 62.8%

Methods & Experiments

Evaluation

Accuracy

Precision

Recall

F1-Score

Predicted

		0	1	
Actual	0	TN	FP Type I error	Specificity = TN/(TN+FP)
	1	FN Type II error	TP	Recall or Sensitivity = TP/(TP+FN)
		Negative Rate =TN/(FN+TN)	Precision = TP/(TP+FP)	

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{F1 - Score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Majority Class

	Accuracy	Precision	Recall	F1-Score
Train	0.5036	1.0	0.5036	0.6699
Development	0.4348	1.0	0.4348	0.6060
Test	0.4348	1.0	0.4348	0.6060

Logistic Regression

Only BERT embeddings of utterance are used.

	Accuracy	Precision	Recall	F1-Score
Train	0.9674	0.9674	0.9674	0.9674
Development	0.5429	0.5143	0.5454	0.5294
Test	0.6470	0.5588	0.6786	0.6129

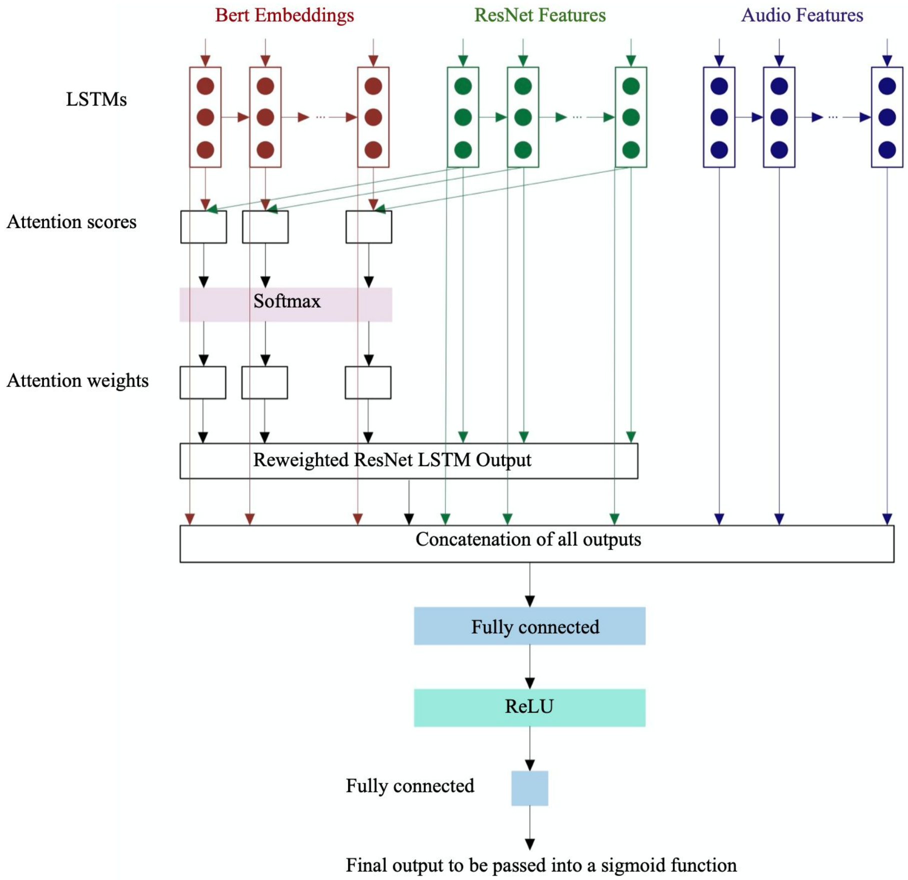
Simple LSTM

Learning Rate: 0.001
Hidden Size: 300
Epochs: 15
Batch size: 32

	Accuracy	Precision	Recall	F1-Score
Train	0.8025	0.8849	0.7616	0.8186
Development	0.7246	0.8108	0.7142	0.7595
Test	0.6232	0.8	0.5455	0.6486

LSTM with Attention

- Unidirectional LSTM for each feature
- Attention scores calculated from textual and visual outputs
- Attention weights applied to visual outputs
- 2 Fully connected resizing layers
- ReLU activation



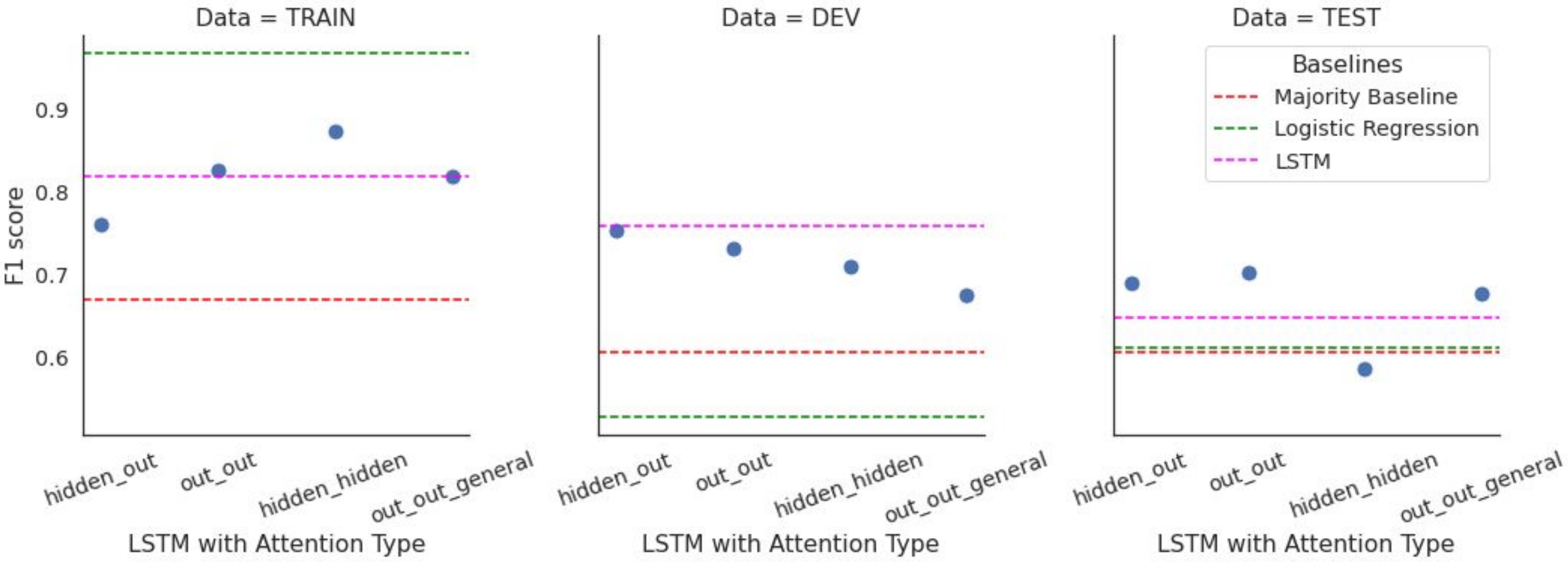
Hyperparameter Tuning

Hyperparameter	Value	Accuracy	Precision	Recall	F1 score
Learning rate	0.01	0.5362	1	0.5362	0.6981
	0.001	0.6811	0.8648	0.653	0.7441
	0.0005	0.6666	0.8648	0.64	0.7356
	0.0001	0.6811	0.5405	0.8	0.6451
Hidden size	100	0.6521	0.8378	0.6326	0.7209
	200	0.6086	0.5405	0.6666	0.597
	300	0.6811	0.8648	0.653	0.7441
	400	0.5797	0.4324	0.6666	0.5245
Batch size	8	0.5797	0.4324	0.6666	0.5245
	16	0.5797	0.3783	0.7	0.4912
	32	0.6811	0.8648	0.653	0.7441
	64	0.6376	0.7567	0.6363	0.6913

LSTM with Attention

	Accuracy	Precision	Recall	F1-score
LSTM without attention	0.6232	0.8	0.5455	0.6486
LSTM hidden_out	0.6232	**0.9667**	0.5370	0.6905
LSTM out_out	0.6812	0.8667	0.5909	**0.7027**
LSTM hidden_hidden	0.6522	0.5667	0.6071	0.5862
out_out_general	**0.7101**	0.7	**0.6563**	0.6774

Comparing all the LSTMs with attention



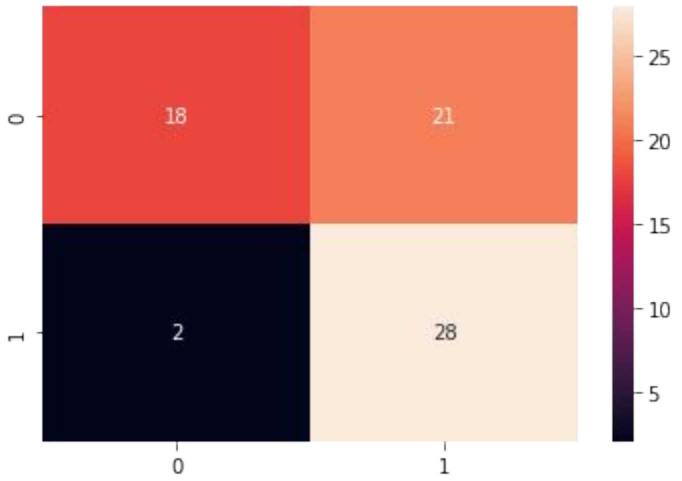
Error Analysis

Quantitative

False Positives > True Negatives

False Negatives << True Positives

Model learns a bias towards predicting data as sarcastic!



Error Analysis

Qualitative

- High pitch mostly detected as sarcastic
- Model learns a bias towards some speakers (like Chandler, Sheldon)
- Misinterprets a joke as a sarcastic comment
- Requires context sometimes, utterance isn't always enough

Conclusion

Implications & Improvements

- Surpasses the model presented in Castro et al. 2019
- In terms of F1 score, as our model is still heavily biased toward sarcasm
- More data entries relative to features
- Inclusion of context related features
- Implementation of more attention mechanisms