Fine-tuning DistilBERT for South Park character classification

Project in TDDE09

Olle Ulfvin, Tobias Grandin Karanta,
Philip Gustafsson and Matteo Cutroni

Content

DATA-SET

BASELINE MODEL

METHODS USED

RESULTS

ANALYSIS

Data-Set

- Consisted of [episode | character | line]
- A lot of characters, we chose top 5 characters
- Imbalanced

Character	Number of Lines (Training)	Number of Lines (Evaluation)
Cartman	6120	1530
Stan	3695	924
Kyle	3481	870
Randy	1897	474
Butters	1435	359
Total	16628	4157

The Task

Baseline Model

- DistilBERT as pre-trained model
- Fine-tuned on South Park data-set
- Learning rate 1e-4, Batch size 8, no weight decay

	Macro F1	Accuracy
Baseline	0.504004	0.528506

CW-Model

- Baseline but with Class Weights
- Introduced to combat class imbalance

	Macro F1	Accuracy
Baseline	0.504004	0.528506
CW	0.499149	0.521289

HPFT-Model

- Hyperparameter fine-tuning
- Large search for best hyperparameters
- Learning rate 5e-5, Batch size 8 and weight decay 0.1

	Macro F1	Accuracy
Baseline	0.504004	0.528506
CW	0.499149	0.521289
HPFT	0.509452	0.532355

LLRD-Model

- Layer-wise learning rate decay
- Lower layer => Lower learning rate
- Starting learning rate 5e-5 and a decay rate of 0.9.

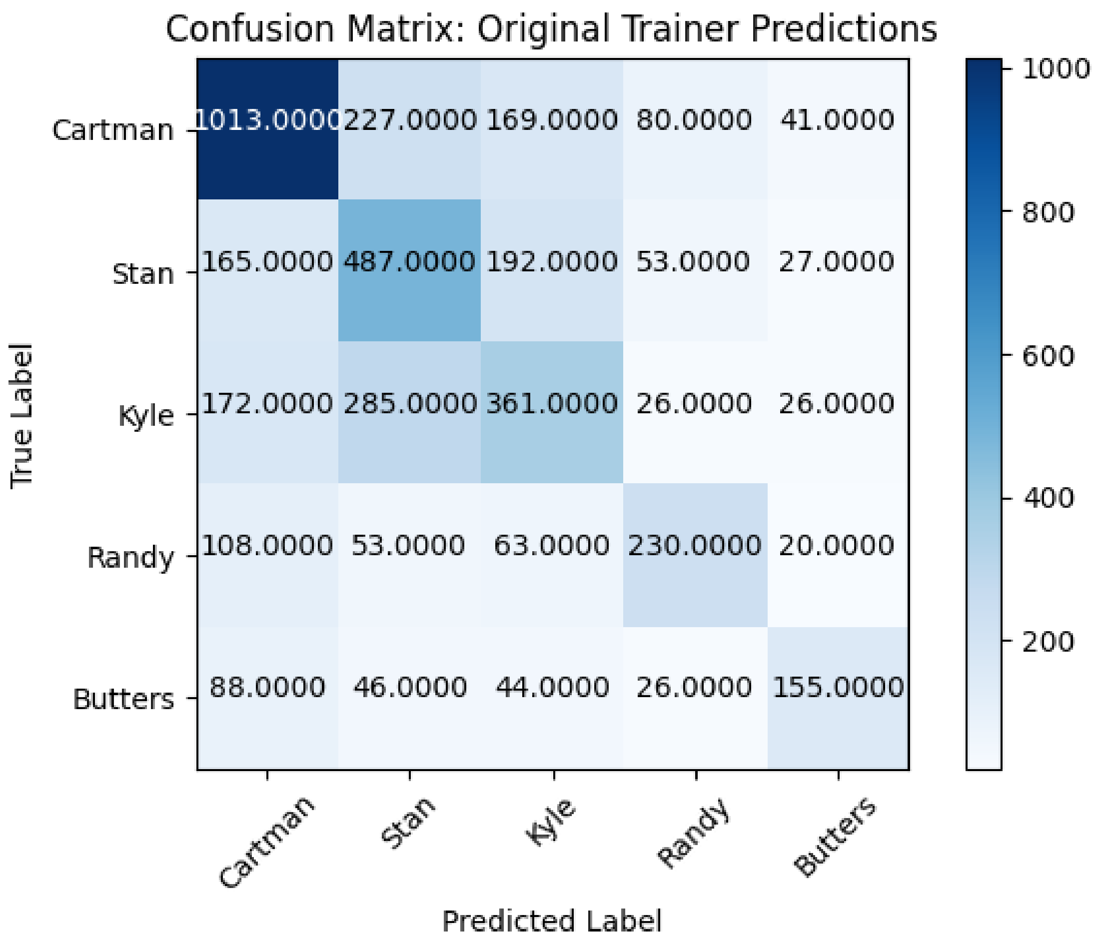
	Macro F1	Accuracy
Baseline	0.504004	0.528506
CW	0.499149	0.521289
HPFT	0.509452	0.532355
LLRD	0.508115	0.528506

Classification with Alternating Normalization (CAN)

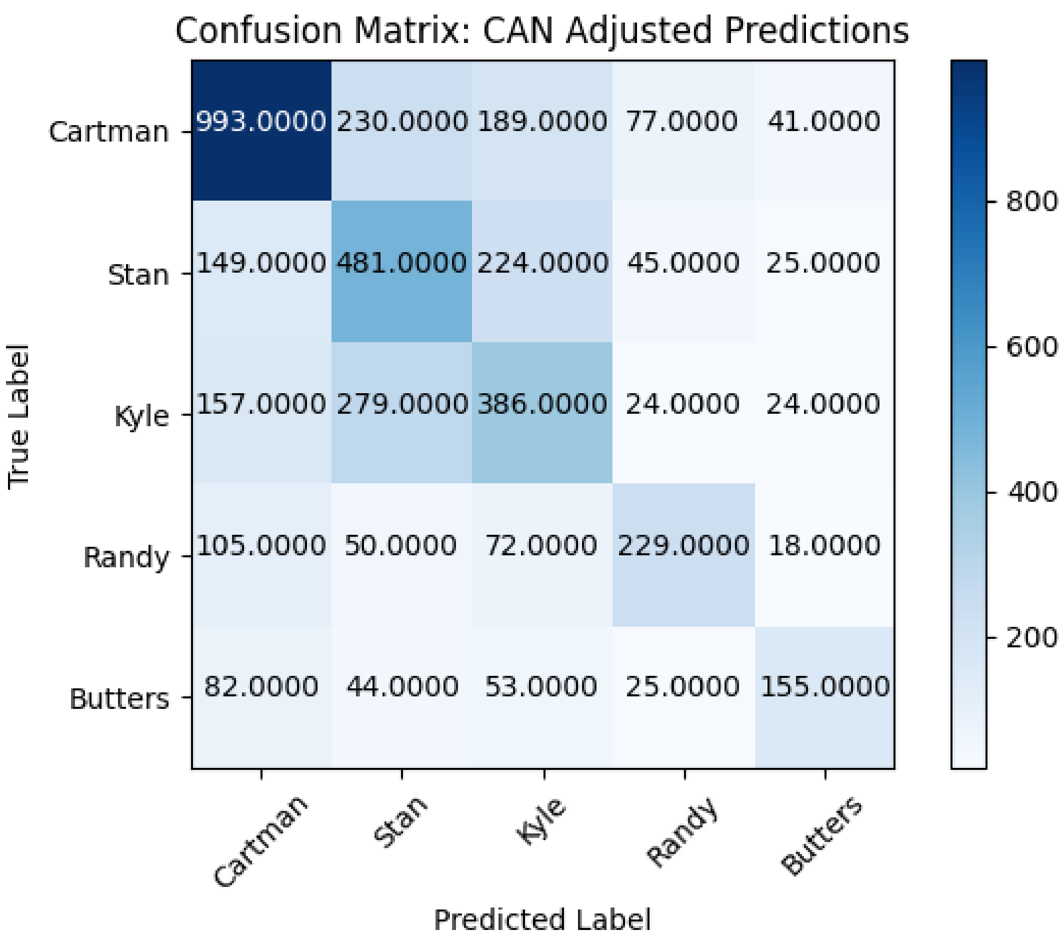
- Non-parametric post processing technique
- Refines predictions for ambiguous examples
- Normalizes probability distributions
- Proven to improve results for classifiers

```
Example 1:
Input text:                 Yeah. Good job, wizard fat ass! Now we're totally lost.
True label:                 Kyle
Original prediction:        Cartman
CAN adjusted prediction:    Kyle
Original percentages:       Cartman: 45.28%, Stan: 13.85%, Kyle: 38.26%, Randy: 0.34%, Butters: 2.27%
CAN adjusted percentages:   Cartman: 46.19%, Stan: 0.67%, Kyle: 53.14%, Randy: 0.00%, Butters: 0.00%
```

CAN Results



Confusion Matrix: Original Trainer Predictions
True Label
Predicted Label
Cartman
Stan
Kyle
Randy
Butters
1013.0000
227.0000
169.0000
80.0000
41.0000
165.0000
487.0000
192.0000
53.0000
27.0000
172.0000
285.0000
361.0000
26.0000
26.0000
108.0000
53.0000
63.0000
230.0000
20.0000
88.0000
46.0000
44.0000
26.0000
155.0000
1000
800
600
400
200



Confusion Matrix: CAN Adjusted Predictions
True Label
Predicted Label
Cartman
Stan
Kyle
Randy
Butters
993.0000
230.0000
189.0000
77.0000
41.0000
149.0000
481.0000
224.0000
45.0000
25.0000
157.0000
279.0000
386.0000
24.0000
24.0000
105.0000
50.0000
72.0000
229.0000
18.0000
82.0000
44.0000
53.0000
25.0000
155.0000
800
600
400
200

Results

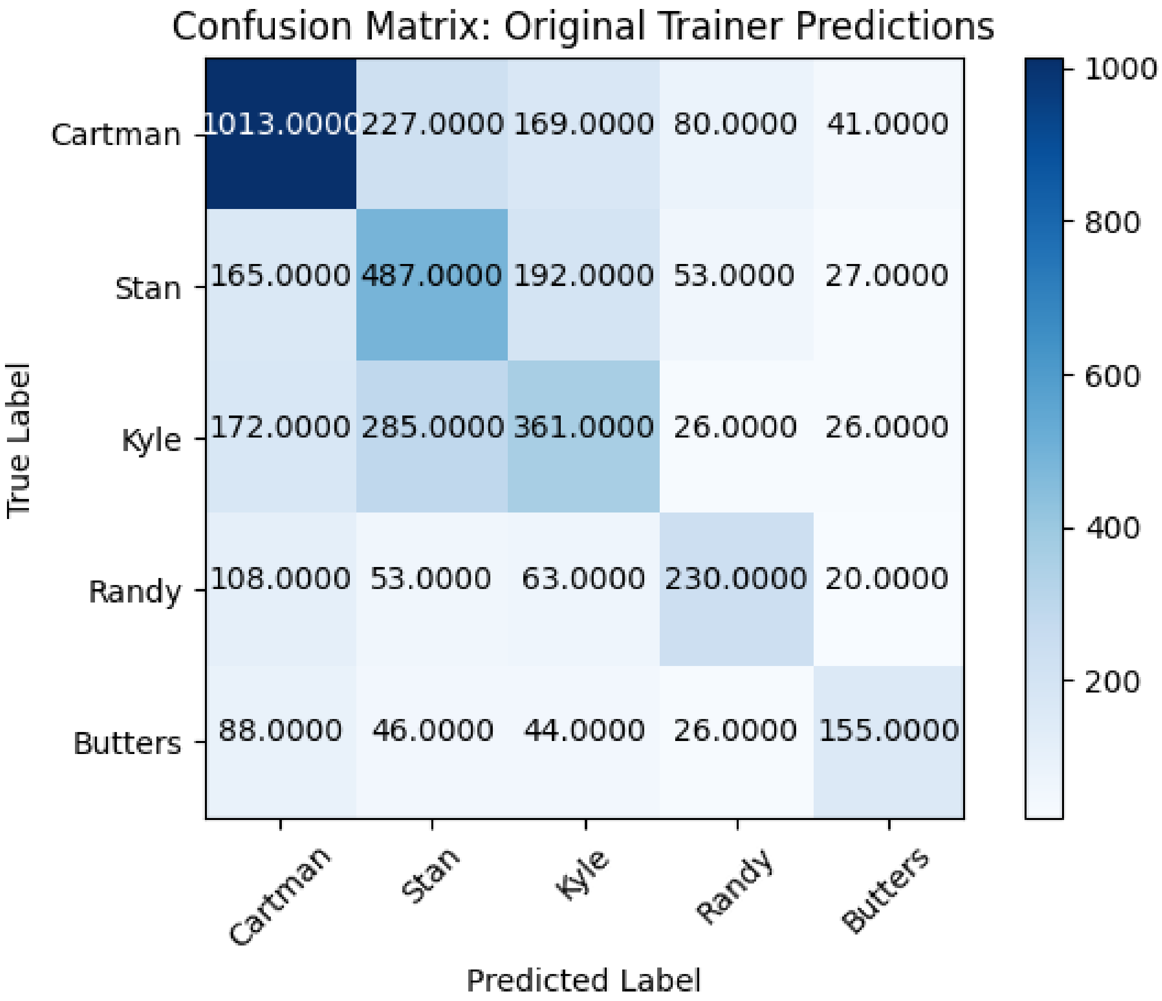
	Macro F1	Accuracy
Baseline	0.504004	0.528506
CW	0.499149	0.521289
HPFT	0.509452	0.532355
LLRD	0.508115	0.528506

Why are the results not improving?

- In more ambiguous cases, it is somewhat uncertain.

Predicted Label	Attribution Label	Word Importance
2 (0.06)	**I can't believe what I'm seeing.**	[CLS] i can ' t believe what i ' m seeing . [SEP]
2 (0.33)	**I can't believe what I'm seeing.**	[CLS] i can ' t believe what i ' m seeing . [SEP]
2 (0.51)	**I can't believe what I'm seeing.**	[CLS] i can ' t believe what i ' m seeing . [SEP]
2 (0.05)	**I can't believe what I'm seeing.**	[CLS] i can ' t believe what i ' m seeing . [SEP]
2 (0.06)	**I can't believe what I'm seeing.**	[CLS] i can ' t believe what i ' m seeing . [SEP]

Confusion matrix



What could the model do?

- Model could connect some words with some characters

Predicted Label	Attribution Label	Word Importance
3 (0.00)	and? your turn, sharon.	[CLS] and ? your turn , sharon . [SEP]
3 (0.00)	and? your turn, sharon.	[CLS] and ? your turn , sharon . [SEP]
3 (0.00)	and? your turn, sharon.	[CLS] and ? your turn , sharon . [SEP]
3 (1.00)	and? your turn, sharon.	[CLS] and ? your turn , sharon . [SEP]
3 (0.00)	and? your turn, sharon.	[CLS] and ? your turn , sharon . [SEP]

Conclusion (results)

Model can find character-specific patterns, but…

Many sentences does not alone contain enough information to deduce the character who said it (see ambiguous cases).


Questions?