Thesis baseline experiment

Chen Wang

June 2020

1 Experiment settings

Hardware:

- Server: Liacs DS Lab, duranium.liacs.nl
- GPU: GeForce GTX 980 Ti

Environment:

- BranchLSTM: Python 2.7, theano, lasage
- DistillBert: Python 3.6. ktrain

2 Results

2.1 BranchLSTM

	Accuracy	**Marco F**	S	D	Q	C
Development	0.782	0.561	0.621	0.000	0.762	0.860
Testing	**0.784**	0.434	0.403	0.000	0.462	0.873

Table 1: Results on the development and testing sets. Accuracy and F1 scores: macro-averaged and per class (S: supporting, D: denying, Q: querying, C: commenting).(Kochkina et al., 2017)

Depth	Tweets	S	D	Q	C	Accuracy	MacroF	S	D	Q	C
0	28	26	2	0	0	0.929	0.481	**0.963**	0.000	0.000	0.000
1	704	61	60	81	502	0.739	0.348	0.000	0.000	**0.550**	**0.842**
2	128	3	6	7	112	0.875	0.233	0.000	0.000	0.000	**0.933**
3	60	2	1	5	52	0.867	0.232	0.000	0.000	0.000	**0.929**
4	41	0	0	3	38	0.927	0.481	0.000	0.000	0.000	**0.962**
5	27	1	0	1	25	0.926	0.321	0.000	0.000	0.000	**0.961**
6+	61	1	2	9	49	0.803	0.223	0.000	0.000	0.000	0.891

Table 2: Number of tweets per depth and performance at each of the depths.(Kochkina et al., 2017)

Label \ **Prediction**	**C**	**D**	**Q**	**S**
Commenting	760	0	12	6
Denying	68	0	1	2
Querying	69	0	36	1
Supporting	67	0	1	26

Table 3: Confusion matrix for testing set predictions.(Kochkina et al., 2017)

2.2 DistillBERT

Epochs	Learning rate	batch size	maximum text length
4	3e-5	6	350

Table 4: DistillBERT hyperparameters

	Accuracy	**Marco F**	S	D	Q	C
Development						
Testing	**0.751**	0.526	0.848	0.250	0.632	0.373

Table 5: DistillBERT Results on the development and testing sets. Accuracy and F1 scores: macro-averaged and per class (S: supporting, D: denying, Q: querying, C: commenting).

Depth	Tweets	S	D	Q	C	Accuracy	MacroF	S	D	Q	C
0	28	26	2	0	0	0.929	0.549	**0.667**	0.000	0.000	0.000
1	704	61	60	81	502	0.720	0.454	0.086	0.265	**0.636**	**0.830**
2	128	3	6	7	112	0.805	0.393	0.000	0.235	0.444	**0.894**
3	60	2	1	5	52	0.850	0.609	0.400	0.400	0.727	**0.909**
4	41	0	0	3	38	0.805	0.435	0.000	0.000	0.000	**0.882**
5	27	1	0	1	25	0.778	0.219	0.000	0.000	0.000	**0.875**
6+	61	1	2	9	49	0.770	0.392	0.000	0.000	0.706	0.863

Table 6: Number of tweets per depth and performance at each of the depths.

Label \ **Prediction**	**C**	**D**	**Q**	**S**
Commenting	675	39	31	33
Denying	49	17	4	1
Querying	33	4	66	3
Supporting	57	5	2	30

Table 7: Confusion matrix for testing set predictions.

3 Comparison

3.1 Comparison between Table5 and Table 1

	Accuracy	**Marco F**	S	D	Q	C
Development						
Testing	**-0.033**	0.092	0.445	0.250	0.170	**-0.500**

Table 8: Comparison Results on the development and testing sets. Accuracy and F1 scores: macro-averaged and per class (S: supporting, D: denying, Q: querying, C: commenting).

3.2 Comparison between Table6 and Table 2

Depth	Tweets	S	D	Q	C	Accuracy	MacroF
0	28	26	2	0	0	0.000	0.068
1	704	61	60	81	502	-0.019	0.106
2	128	3	6	7	112	-0.070	0.160
3	60	2	1	5	52	-0.017	0.377
4	41	0	0	3	38	-0.122	-0.046
5	27	1	0	1	25	-0.148	-0.102
6+	61	1	2	9	49	-0.033	0.169

Table 9: Number of tweets per depth and performance comparison at each of the depths.

3.3 Comparison between Table7 and Table 3

Label \ Prediction	C	D	Q	S
Commenting	-85	39	19	27
Denying	-19	17	3	-1
Querying	-36	4	30	2
Supporting	-10	5	1	4

Table 10: Comparison Confusion matrix for testing set predictions.

References

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*, 2017.