A Supplemental Material

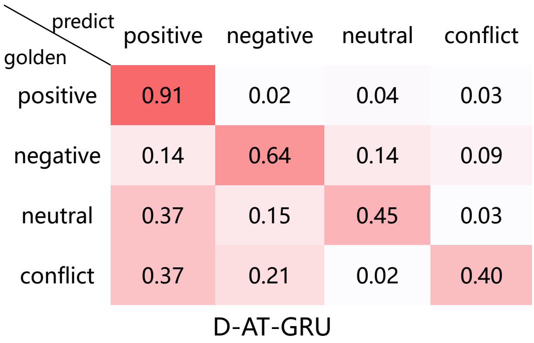
A.1 Impact of Conflict Opinion

Table 4 shows the impact of adding extra types of sentiment (*neutral* in 3-way, *conflict* in 4-way) when train and test the model. In particular, when we add *conflict* sentiment, when accuracy drop is large (6%), considering only 5.1% of the dataset is *conflict*. This indicates that AT-LSTM model has difficulty in dealing with *conflict* opinions.

	2-way	**3-way**	**4-way**
AT-LSTM	89.5%	83.1%	77.1%

Table 4: The accuracy of AT-LSTM doing sentiment classification of 2-way (*positive*, *negative*), 3-way (*positive*, *negative*, *neutral*), and 4-way (*positive*, *negative*, *neutral*, *conflict*).

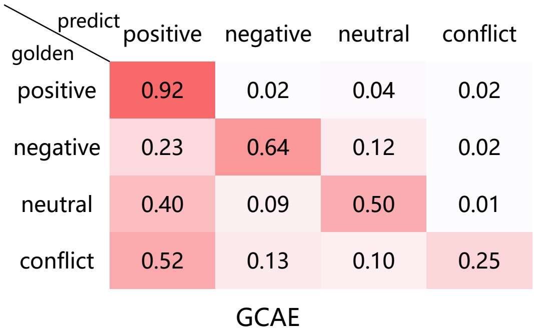
A.2 Confusion Matrix



predict golden	positive	negative	neutral	conflict
positive	0.91	0.02	0.04	0.03
negative	0.14	0.64	0.14	0.09
neutral	0.37	0.15	0.45	0.03
conflict	0.37	0.21	0.02	0.40

D-AT-GRU

Figure 3: The normalized confusion matrixe of D-AT-LSTM.

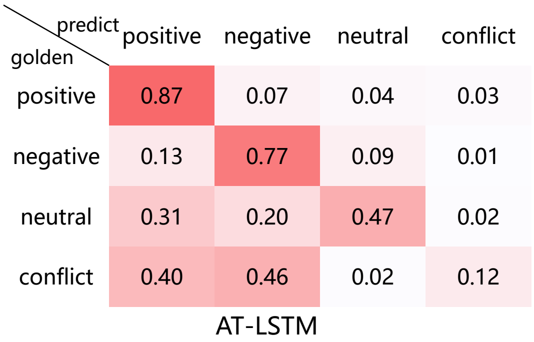


predict golden	positive	negative	neutral	conflict
positive	0.92	0.02	0.04	0.02
negative	0.23	0.64	0.12	0.02
neutral	0.40	0.09	0.50	0.01
conflict	0.52	0.13	0.10	0.25

GCAE

Figure 4: The normalized confusion matrixe of GCAE.

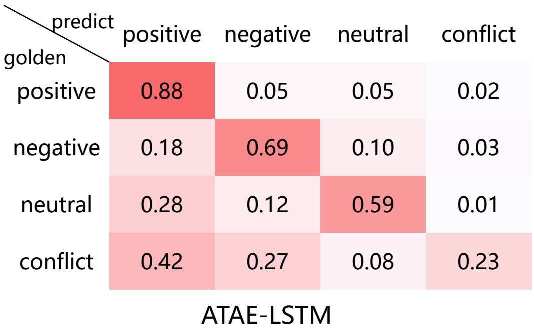
Figure 3, 4, 6, 5 shows the confusion matrixes of models we test in experiment. Our proposed D-AT-GRU model achieve the highest accuracy on *conflict* category.



predict golden	positive	negative	neutral	conflict
positive	0.87	0.07	0.04	0.03
negative	0.13	0.77	0.09	0.01
neutral	0.31	0.20	0.47	0.02
conflict	0.40	0.46	0.02	0.12

AT-LSTM

Figure 5: The normalized confusion matrixe of AT-LSTM.



predict golden	positive	negative	neutral	conflict
positive	0.88	0.05	0.05	0.02
negative	0.18	0.69	0.10	0.03
neutral	0.28	0.12	0.59	0.01
conflict	0.42	0.27	0.08	0.23

ATAE-LSTM

Figure 6: The normalized confusion matrixe of ATAE-LSTM.

A.3 Weighted Loss

We try using weighted loss function to give *conflict* samples more weight during training. Through this way, the accuracy on *conflict* test samples can reach 55.77%. However, the accuracy on other classes decrease significantly, which make the overall accuracy be 74.39%. Thus, simply adjusting the weight of *conflict* samples cannot fix the problem of data sparsity.

A.4 D-ATAE-GRU

We also try to concatenate word embeddings with aspect embeddings as the inputs of GRU (D-ATAE-GRU) similar with ATAE-LSTM. However, its improvement is minor (about 0.1%). Considering the amount of extra parameters it adds to our model (Fig. 5), we choose not to integrate the idea of ATAE-LSTM into our model.

A.5 Multi-Task Learning Perspective

The proposed D-AT-GRU model can be understood from a multi-task learning perspective. One task is to determine whether there is *positive* sentiment expressed towards given aspect. The other task is to determine whether there is *negative* sen-

Model	Parameters
AT-LSTM	1082.4k
ATAE-LSTM	1442.4k
GCAE	751.6k
B-AT-GRU	903.9k
B-ATAE-GRU	1173.9k

Table 5: The amount of parameters, ignoring word embeddings and aspect embeddings.

timent expressed towards given aspect. These two tasks are related since they both need encoded text features to further analyse sentiment. The difference is which type of words they need to attend to. Thus the embedding layers and GRU layer are share parameters, while the attention layers and classification layers are independent. Through this way, the two tasks learn and share the same text features, but select different regions they attend to. In addition, we add orthogonal regularization to maintain the diversity of the two attentions.