Reproduction of TIMME: Twitter ideology-detection via multi-task multi-relational embedding

4268110 Eunji An

1. Paper to be replicated:

Xiao, Z., Song, W., Xu, H., Ren, Z., & Sun, Y. (2020, August). TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2258-2268).

(Github page) https://github.com/PatriciaXiao/TIMME

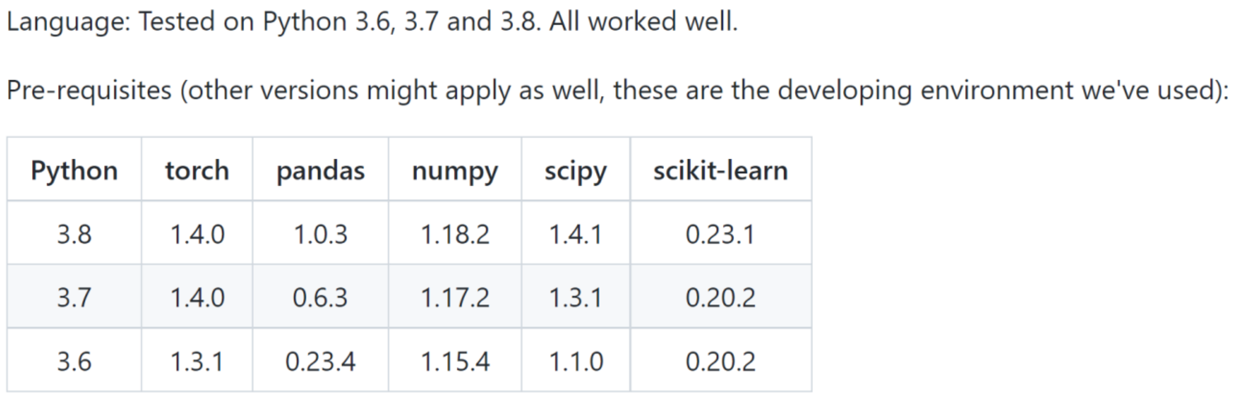
(Paper) https://arxiv.org/abs/2006.01321

2. Reason of selection:

Since my bachelor thesis is about ideology detection, I thought it would be nice to replicate one paper regarding the same topic. I found this paper and its Github webpage. In the Github page, the authors uploaded the whole code and the datasets they used in the paper.

3. Settings:

The environment settings from authors are like below.



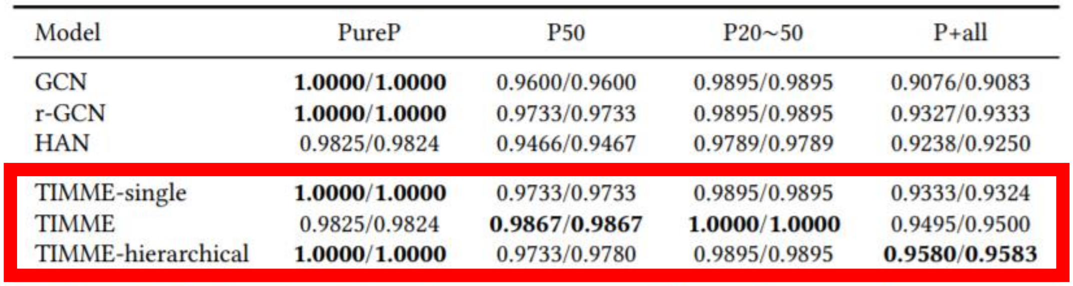
Language: Tested on Python 3.6, 3.7 and 3.8. All worked well.

Pre-requisites (other versions might apply as well, these are the developing environment we've used):

Python	torch	pandas	numpy	scipy	scikit-learn
3.8	1.4.0	1.0.3	1.18.2	1.4.1	0.23.1
3.7	1.4.0	0.6.3	1.17.2	1.3.1	0.20.2
3.6	1.3.1	0.23.4	1.15.4	1.1.0	0.20.2

And my setting was Python 3.8, torch 1.9.0, pandas 1.1.3, numpy 1.19.2, scipy 1.5.2, scikit-learn 0.23.2

4. Comparison of the test result:



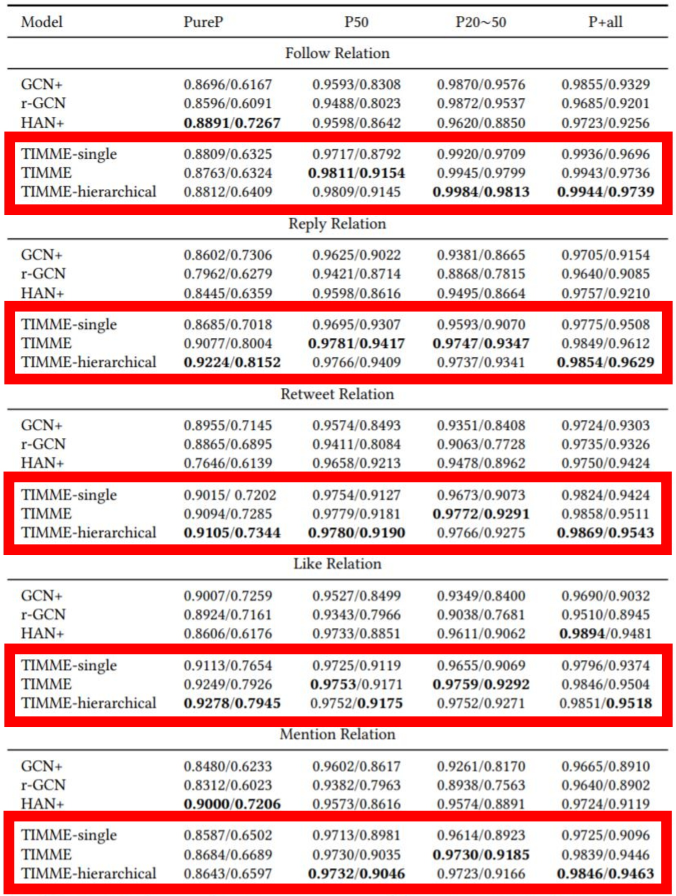
Model	PureP	P50	P20~50	P+all
GCN	**1.0000/1.0000**	0.9600/0.9600	0.9895/0.9895	0.9076/0.9083
r-GCN	**1.0000/1.0000**	0.9733/0.9733	0.9895/0.9895	0.9327/0.9333
HAN	0.9825/0.9824	0.9466/0.9467	0.9789/0.9789	0.9238/0.9250
TIMME-single	**1.0000/1.0000**	0.9733/0.9733	0.9895/0.9895	0.9333/0.9324
TIMME	0.9825/0.9824	**0.9867/0.9867**	**1.0000/1.0000**	0.9495/0.9500
TIMME-hierarchical	**1.0000/1.0000**	0.9733/0.9780	0.9895/0.9895	**0.9580/0.9583**

Table 2: Node classification measured by F1-score/accuracy.

This is the table of test result that I want to replicate. I would like to use TIMME models only, since this paper is about TIMME. And I will also use four datasets for comparison. Even though there is another table of result, called Table 3: Link-prediction measured by ROC-AUC/PR-AUC in this paper, I would like to focus on F1-score and accuracy of simple node classification of TIMME models.

Model	PureP	P50	P20~50	P+all
N/A (Classification)	0.9825/ 0.9825	0.9064/ 0.9067	0.9892/ 0.9895	0.9155/ 0.9167
TIMME-single	0.3596/ 0.5614 (epoch 600)	0.3391/ 0.5132 (epoch 300)	0.3533/ 0.5464 (epoch 200)	0.3698/ 0.5868 (epoch 200)
TIMME	1.0000/ 1.0000	0.9302/ 0.9333	0.9444/ 0.9474	Not possible (not enough memory)
TIMME-hierarchical	0.9824/ 0.9825	0.9600/ 0.9600	0.9682/ 0.9684	Not possible (not enough memory)

Table 1: Node classification measured by F1-score/ accuracy (Other than TIMME-single model, all of them have epoch 20).



Model	PureP	P50	P20~50	P+all
		Follow Relation		
GCN+	0.8696/0.6167	0.9593/0.8308	0.9870/0.9576	0.9855/0.9329
r-GCN	0.8596/0.6091	0.9488/0.8023	0.9872/0.9537	0.9685/0.9201
HAN+	**0.8891/0.7267**	0.9598/0.8642	0.9620/0.8850	0.9723/0.9256
TIMME-single	0.8809/0.6325	0.9717/0.8792	0.9920/0.9709	0.9936/0.9696
TIMME	0.8763/0.6324	**0.9811/0.9154**	0.9945/0.9799	0.9943/0.9736
TIMME-hierarchical	0.8812/0.6409	0.9809/0.9145	**0.9984/0.9813**	**0.9944/0.9739**
		Reply Relation		
GCN+	0.8602/0.7306	0.9625/0.9022	0.9381/0.8665	0.9705/0.9154
r-GCN	0.7962/0.6279	0.9421/0.8714	0.8868/0.7815	0.9640/0.9085
HAN+	0.8445/0.6359	0.9598/0.8616	0.9495/0.8664	0.9757/0.9210
TIMME-single	0.8685/0.7018	0.9695/0.9307	0.9593/0.9070	0.9775/0.9508
TIMME	0.9077/0.8004	**0.9781/0.9417**	**0.9747/0.9347**	0.9849/0.9612
TIMME-hierarchical	**0.9224/0.8152**	0.9766/0.9409	0.9737/0.9341	**0.9854/0.9629**
		Retweet Relation		
GCN+	0.8955/0.7145	0.9574/0.8493	0.9351/0.8408	0.9724/0.9303
r-GCN	0.8865/0.6895	0.9411/0.8084	0.9063/0.7728	0.9735/0.9326
HAN+	0.7646/0.6139	0.9658/0.9213	0.9478/0.8962	0.9750/0.9424
TIMME-single	0.9015/ 0.7202	0.9754/0.9127	0.9673/0.9073	0.9824/0.9424
TIMME	0.9094/0.7285	0.9779/0.9181	**0.9772/0.9291**	0.9858/0.9511
TIMME-hierarchical	**0.9105/0.7344**	**0.9780/0.9190**	0.9766/0.9275	**0.9869/0.9543**
		Like Relation		
GCN+	0.9007/0.7259	0.9527/0.8499	0.9349/0.8400	0.9690/0.9032
r-GCN	0.8924/0.7161	0.9343/0.7966	0.9038/0.7681	0.9510/0.8945
HAN+	0.8606/0.6176	0.9733/0.8851	0.9611/0.9062	**0.9894**/0.9481
TIMME-single	0.9113/0.7654	0.9725/0.9119	0.9655/0.9069	0.9796/0.9374
TIMME	0.9249/0.7926	**0.9753**/0.9171	**0.9759/0.9292**	0.9846/0.9504
TIMME-hierarchical	**0.9278/0.7945**	0.9752/**0.9175**	0.9752/0.9271	0.9851/**0.9518**
		Mention Relation		
GCN+	0.8480/0.6233	0.9602/0.8617	0.9261/0.8170	0.9665/0.8910
r-GCN	0.8312/0.6023	0.9382/0.7963	0.8938/0.7563	0.9640/0.8902
HAN+	**0.9000/0.7206**	0.9573/0.8616	0.9574/0.8891	0.9724/0.9119
TIMME-single	0.8587/0.6502	0.9713/0.8981	0.9614/0.8923	0.9725/0.9096
TIMME	0.8684/0.6689	0.9730/0.9035	**0.9730/0.9185**	0.9839/0.9446
TIMME-hierarchical	0.8643/0.6597	**0.9732/0.9046**	0.9723/0.9166	**0.9846/0.9463**

Table 3: Link-prediction measured by ROC-AUC/PR-AUC.

	PureP	P50	P20~50	P+all
		Follow Relation		
TIMME-single	0.4628/ 0.2440	0.7827/ 0.4331	0.4424/ 0.2144	0.5942/ 0.2934
TIMME	0.8652/ 0.6200	0.9759/ 0.8957	0.9936/ 0.9774	-
TIMME-hierarchical	0.8781/ 0.6474	0.9758/ 0.8953	0.9933/ 0.9747	-

Reply Relation				
TIMME-single	0.4628/ 0.2325	0.2750/ 0.1674	0.6938/ 0.3727	0.5265/ 0.2845
TIMME	0.8341/0.6258	0.9731/ 0.9322	0.9533/ 0.8939	-
TIMME-hierarchical	0.8210/ 0.6665	0.9747/ 0.9370	0.9548/ 0.8966	-
Retweet Relation				
TIMME-single	0.9031/ 0.7313	0.9769/ 0.9167	0.9642/ 0.9002	0.9767/ 0.9223
TIMME	0.9026/ 0.7247	0.9750/ 0.9122	0.9629/ 0.8960	-
TIMME-hierarchical	0.8975/ 0.7243	0.9750/ 0.9114	0.9639/ 0.8996	-
Like Relation				
TIMME-single	0.5340/ 0.2603	0.4573/ 0.2331	0.7000/ 0.3481	0.7549/ 0.4777
TIMME	0.9049/ 0.7331	0.9737/ 0.9111	0.9605/ 0.8945	-
TIMME-hierarchical	0.9091/ 0.7531	0.9731/ 0.9126	0.9614/ 0.8951	-
Mention Relation				
TIMME-single	0.4666/ 0.2355	0.3432/ 0.1836	0.7089/ 0.4566	0.3885/ 0.1918
TIMME	0.8486/ 0.6354	0.9701/ 0.8916	0.9551/ 0.8780	-
TIMME-hierarchical	0.8512/ 0.6379	0.9698/ 0.8914	0.9561/ 0.8792	-

Table 2: Link-prediction measured by ROC-AUC/PR-AUC.

<Comparison between mine and the authors' results>

- Unlike the authors of the paper, my baseline would be the classification task without any model. I added its F1 score and accuracy in Table1 and unlike the result from any of the authors' models, the result with P50 dataset has the lowest F1-score and accuracy. However, when I look at the authors' results, they always have the lowest F1-score and accuracy in the case of P+all dataset.
- I have the highest F1-score and accuracy with PureP dataset in all of three models, but the authors have the highest score with P20~50 in TIMME model (other than that, the authors also have the highest F1 score and accuracy in PureP dataset).
- Because of the lack of memory with my laptop, I could not manage to run the models for P+all dataset in TIMME and TIMME-hierarchical model. Therefore, I could not get the results of them. (Error message such as "RuntimeError: [enforce fail at ..\c10\core\CPUAllocator.cpp:79] data. DefaultCPUAllocator: not enough memory: you tried to allocate 485248000 bytes")
- While I use the model TIMME-single, I had another error message that I do not have enough memory for further training. I needed to reduce the epoch for the datasets except PureP. So I set epoch 300 for P50 and epoch 200 for P 20~50 and P+all.
- In all of the datasets, I have significantly low f1-score and accuracy when I use TIMME-single model as opposed to the authors' high scores. Even though I set the hyperparameter epoch 600 and single_relation as 0 as the authors did, still my results were low. Maybe there are some other hyperparameters effects to the prediction of TIMME-single model and I did not set it right. However, because I have my laptop's memory issue and it took so long to run the TIMME-single model with 600 epoch, I could not manage to experiment all of the hyperparameter setting to get the high F1 score and accuracy as the authors did.
- Not only F1 score and accuracy but also ROC-AUC and PR-AUC results were, overall, much lower than the authors' when I useTIMME-single model. Only retweet relation results seems appropriate comparing to the authors'.
- One similar point between mine and the authors' ROC-AUC and PR-AUC results is that they also have relatively low results when they use PureP dataset regardless of the type of model. And in my Table2, most of the time I also had lowest ROC-AUC and PR-AUC results in all of the models.