NTULM: Enriching Social Media Text Representations with Non-Textual Units

Jinning Li[1^*], Shubhanshu Mishra[2^], Ahmed El-Kishky[2], Sneha Mehta[2], Vivek Kulkarni[2]
[1]University of Illinois at Urbana-Champaign, [2]Twitter, Inc.,
^Equal Contribution
*Work done during internship at Twitter, Inc.

2022 The 8th Workshop on Noisy User-generated Text (W-NUT)
Oct 16, 2022 — collocated with COLING 2022.

Motivation: Non-Textual Units

Non-Textual Units (NTUs) are the social contexts which appear alongside a social media post, e.g. *Hashtag*, *URL*, *author*, *user mentions* and *media*

https://twitter.com/KQEDscience/status/1528833154492026880

Challenge: Existing models and NTUs

NTUs embedded in the text are broken up by tokenizers diminishing their signal.

Non embedded NTUs are not included.

NTUs have a global context outside of the text.

```
[happy, [UNK], #, world, ##tur, ##tled, ##ay, [UNK],
from, #, deep, ##lo, ##ok, !, let, , s, #, shell,
##ab, ##rate, !, watch, these, crazy, cute, baby,
turtles, take, their, lake, back, in, this, video,
from, our, archives, featuring, conservation,
efforts, by, @, oak, ##zoo, @, sf, ##zoo, and, @,
pre, ##si, ##dio, ##sf, ., http, :, /, /, bit, ., l,
##y, /, y, ##tt, ##urt, ##les]
(Result from tokenizer of bert-base-uncased)
```

https://twitter.com/KQEDscience/status/1528833154492026880

Intuition: Our approach for Non-Textual Units

Inject average NTU embeddings into the Transformer alongside token embeddings.

Pre-compute NTU embeddings using heterogeneous networks, e.g. social engagements for users and Hashtags

```
[happy, [UNK], #, world, ##tur, ##tled, ##ay, [UNK], from, #,
deep, ##lo, ##ok, !, let, , s, #, shell, ##ab, ##rate, !, watch,
these, crazy, cute, baby, turtles, take, their, lake, back, in,
this, video, from, our, archives, featuring, conservation,
efforts, by, @, oak, ##zoo, @, sf, ##zoo, and, @, pre, ##si,
##dio, ##sf, ., http, :, /, /, bit, ., l, ##y, /, y, ##tt,
##urt, ##les] + [@KQEDscience, #WorldTurtleDay, #DeepLook,
#shellabrate, @oakzoo, @sfzoo, @presidiosf, bit.ly/YTTurtles,
Media 1]
```

NTULM Framework

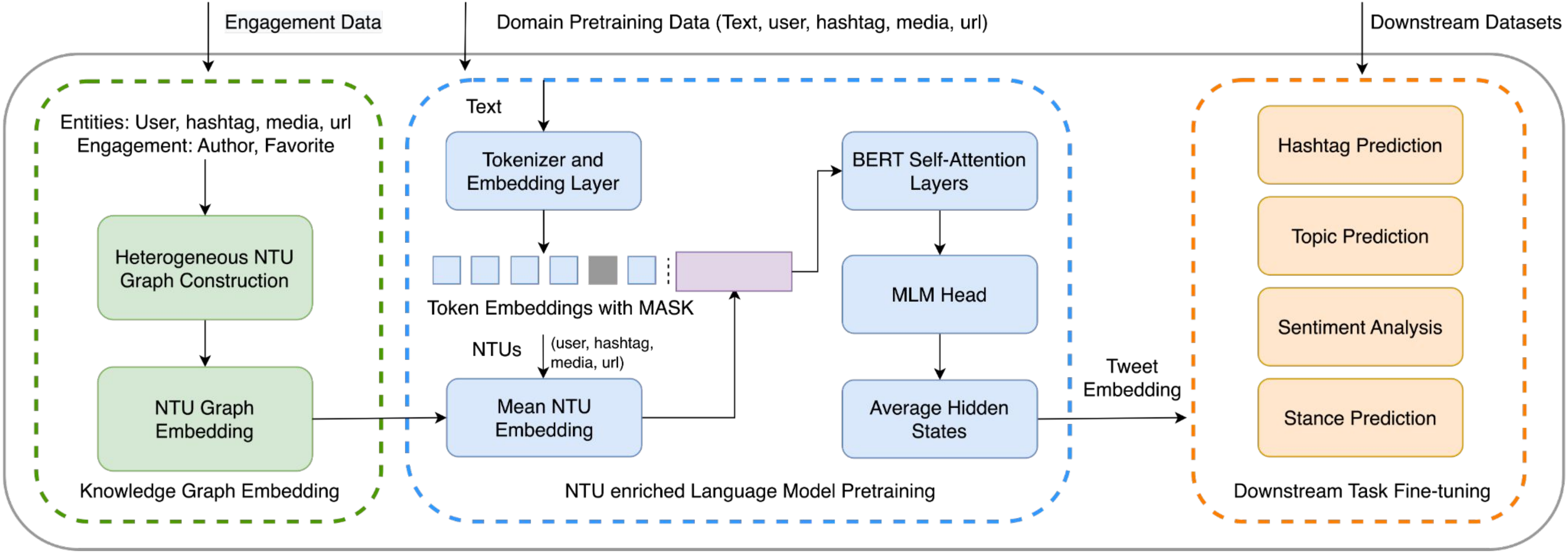


Fig 1: Framework of NTULM

Knowledge Graph Embedding

- **Graph nodes**: author, Hashtag
- **Graph edges**: connect user-Hashtag if user authors, favorites, or is co-mentioned with a Hashtag
- **Training**: TwHIN framework (El-Kishky et al)

Author: *user1*
Tweet: Our paper was accepted at *@WNUT* with *@user2* *@user3* *#nlproc* *#socialmedia*
Favorited by: *user4*, *user5*

Table 1: Example tweet with engagement data of author, mentions, Hashtags, and favorites

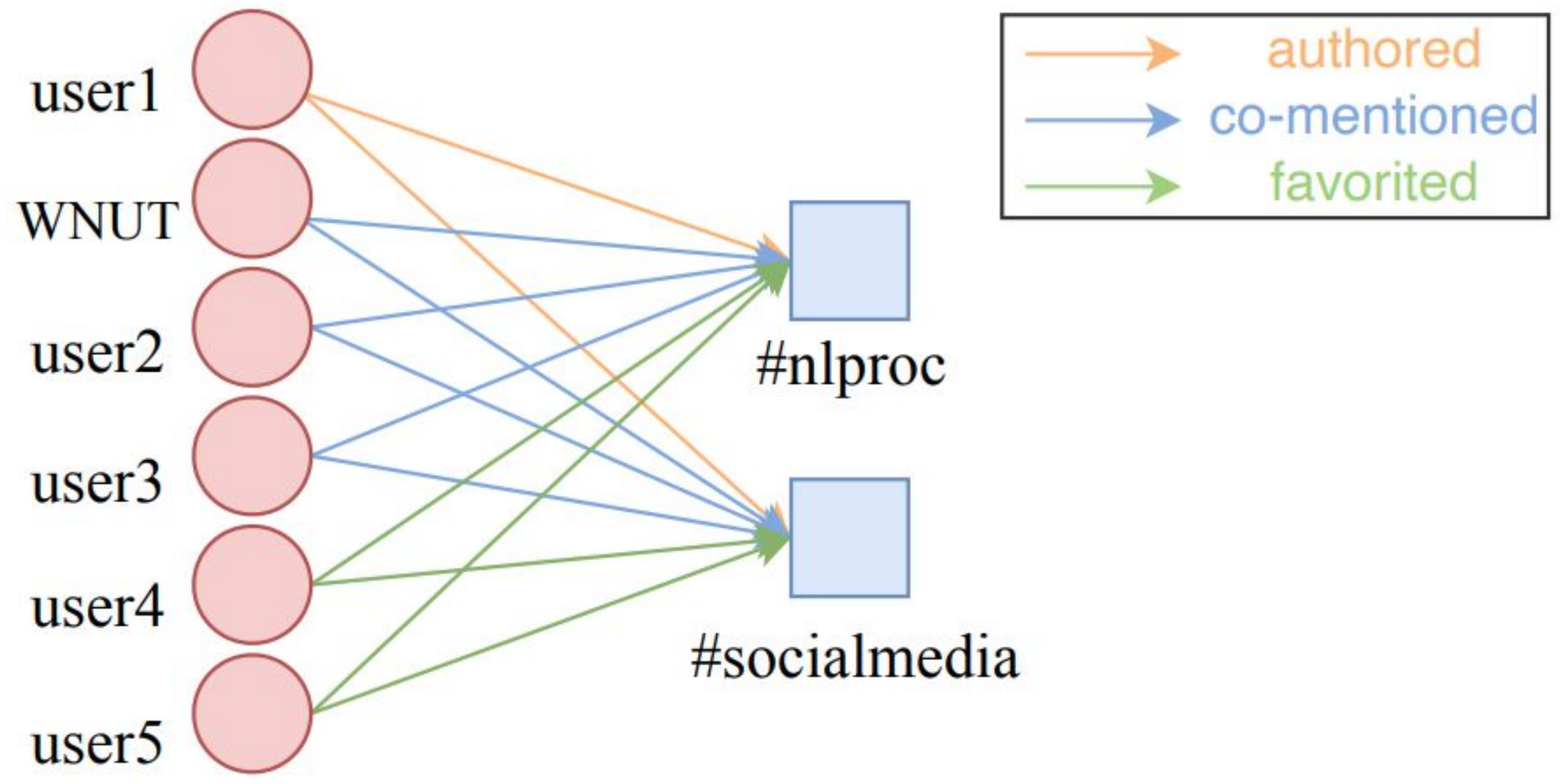
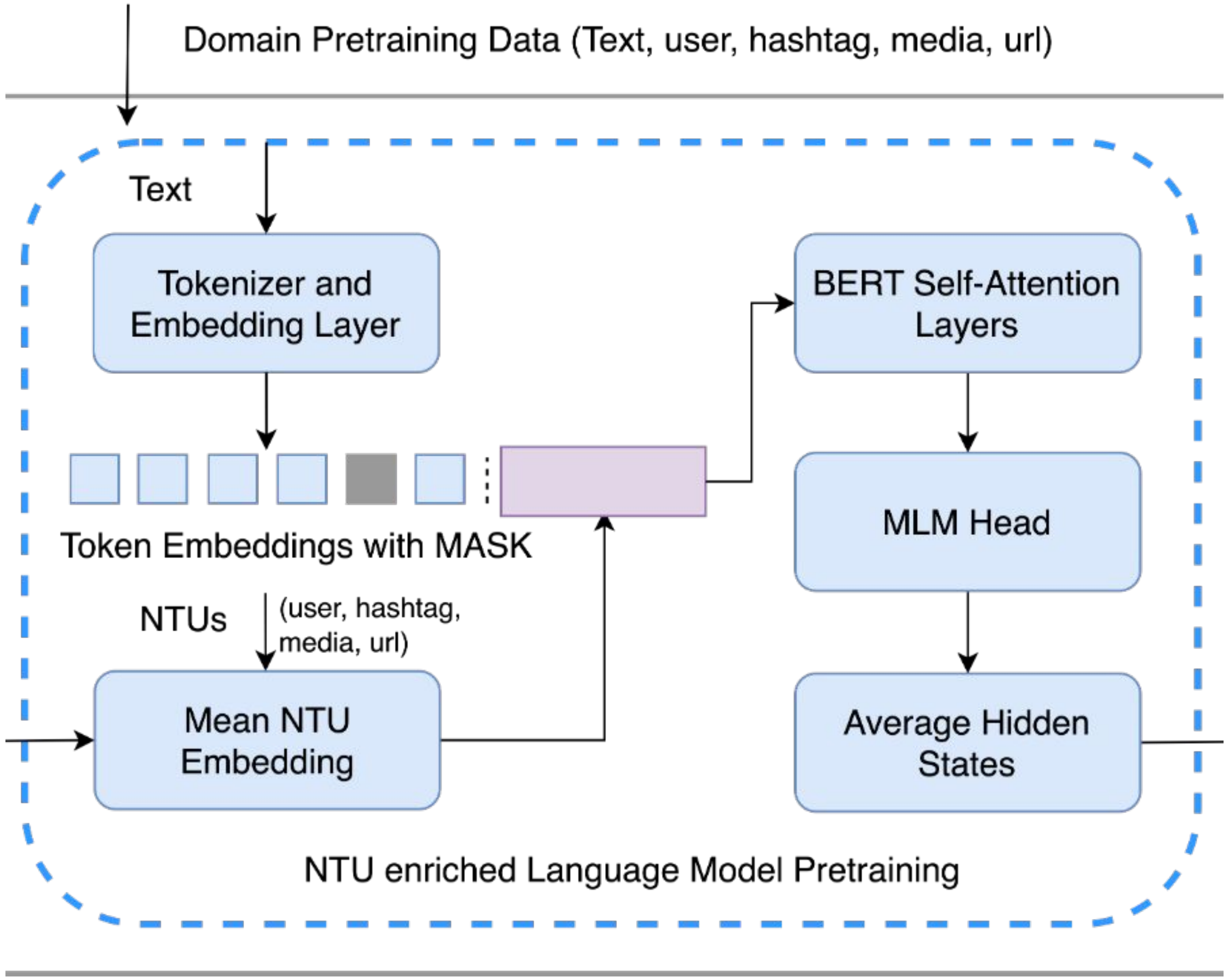


Figure 2: Graph construction with the example data in Table 1 for training NTULM user-Hashtag embeddings.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 2842–2850. https://doi.org/10.1145/3534678.3539080

NTULM: Masked Language Modeling

- Tweet with NTUs, use average NTU embeddings
- Linear projection to map the average NTU embedding from graph space to LM space
- Concatenate NTU embedding to token embeddings
- Average embedding of NTU type for OOV NTUs
- Fine-tune NTULM via MLM



Experiments - Dataset

NTU heterogeneous network: Tweets (2018-01-01~2022-07-01) with Hashtags and their engagements with users, consisting of 60M Hashtags, 255M users, 5B authorship edges, 3B favorite edges, and 0.9B co-mention edges. We only considered users with 10 - 100 unique Hashtags interactions

MLM fine tuning: 1M Tweets sampled from (2022-06-01~2022-06-15).
We also fine-tune BERT without NTUs on these Tweets.

Downstream Tasks: TweetEval, SemEval, SocialMediaIE, Hashtag Pred, Topic

Results: Masked Language Modeling

Model	NTUs	Perplexity bits
BERT	-	4.425
NTULM	author	4.412
NTULM	Hashtag	4.391
NTULM	author+Hashtag	**4.344**

Incorporating NTU embedding improves perplexity

Hashtag embedding is more effective than user embedding, combination is best

Evaluation on Downstream Tasks

Tweet embedding = average final layer hidden states of valid tokens (and NTUs)

Compute all the Tweet embeddings in Downstream Train and Test sets

Train a 2-Layer MLP classifier for downstream tasks using Tweet embeddings

Evaluate using task specific metrics (F1 score, precision, AUC)

Results: All tasks

Model	NTUs	Perplexity bits	Topic MAP	TweetEval mean F1	SemEval 1 mean F1	SemEval 2 mean F1	Hashtag Recall@10	SMIE mean F1
BERT	-	4.425	0.327	0.577	0.527	0.515	0.689	0.548
NTULM	**author**	4.412	0.325	0.579	0.527	**0.548**	0.693	0.548
NTULM	**Hashtag**	4.391	0.339	0.586	0.534	0.545	0.711	0.539
NTULM	**author+Hashtag**	**4.344**	**0.343**	**0.590**	**0.534**	0.545	**0.720**	**0.549**

Incorporating NTU embedding improves downstream task performance

Hashtag embedding is more effective than user embedding, combination is best

NTU Overlap in downstream datasets

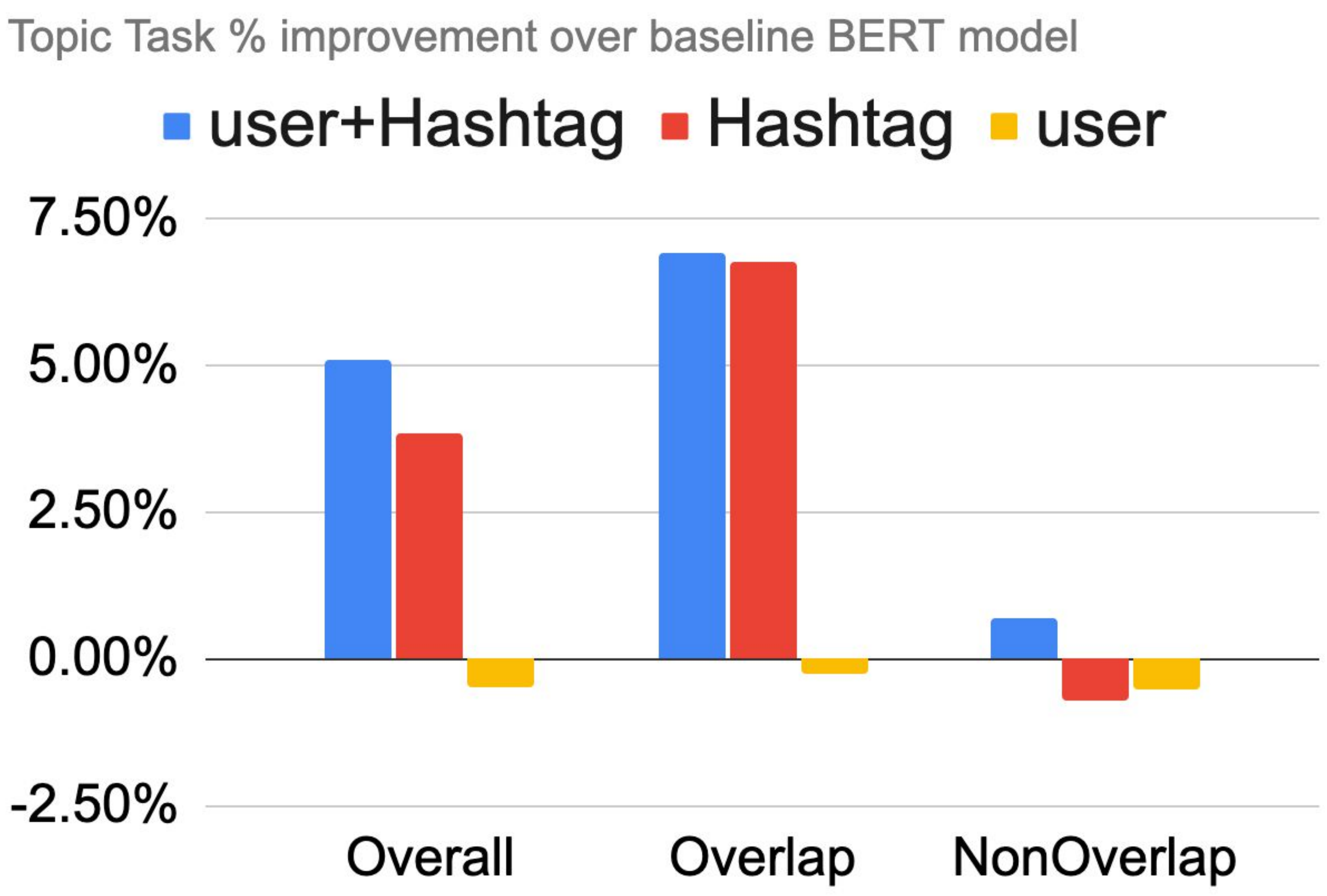
Dataset	Hashtag overlap	User overlap
Hashtag	99%	10%
SemEval	92%	21%
Social Media IE	95%	22%
Topic	99%	14%
TweetEval	98%	0%
Grand Total	95%	14%

Downstream Hashtags more likely to overlap with NTU embeddings than users.

Why is NTULM effective?

Hypothesis:

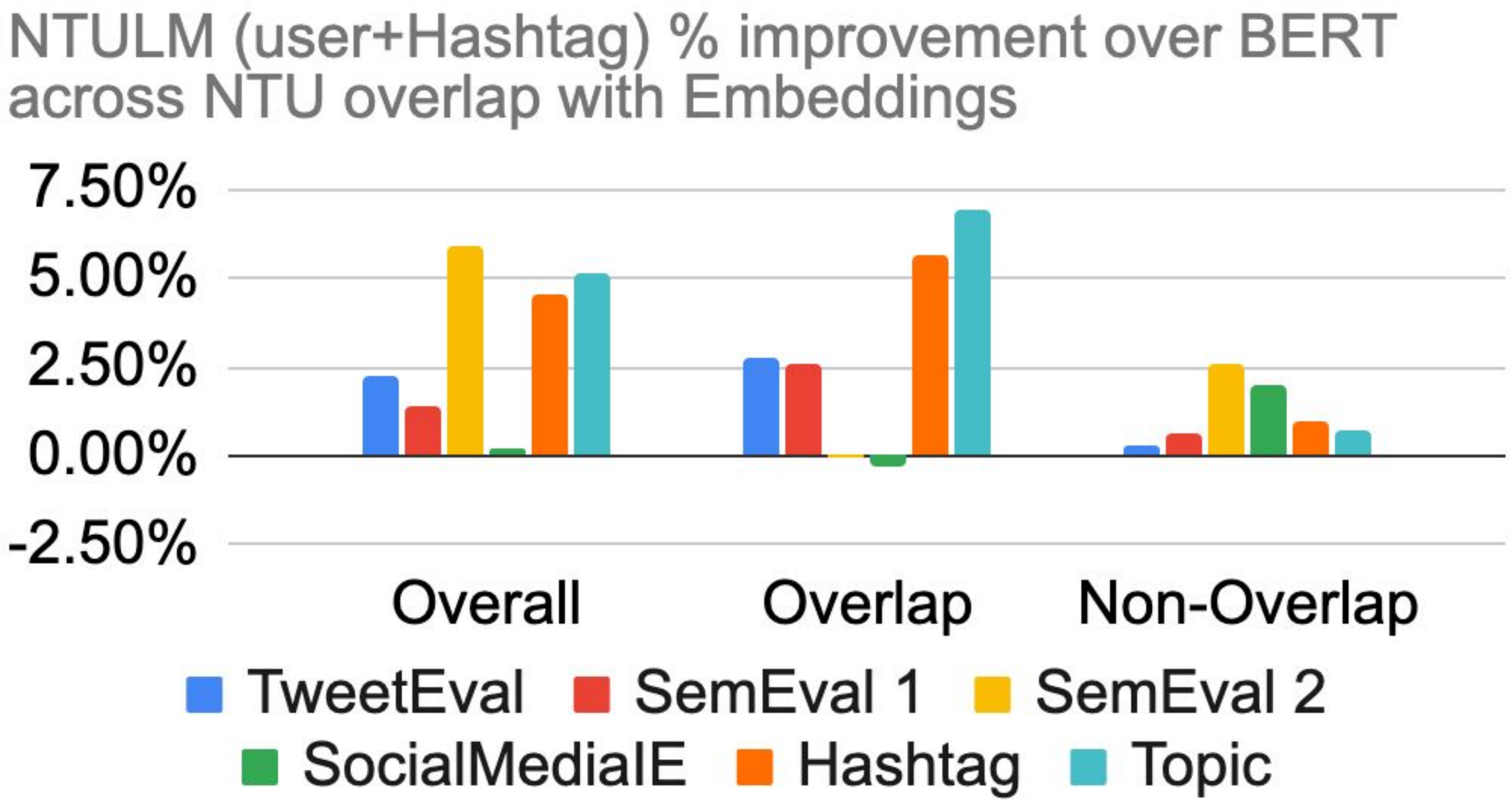
- If NTU is available, NTULM should help.
- If NTU is absent, NTULM should be similar to BERT.



Observation:

- Hypothesis holds
- Gains with Hashtag NTU are much better than user.

Results: Overlap performance

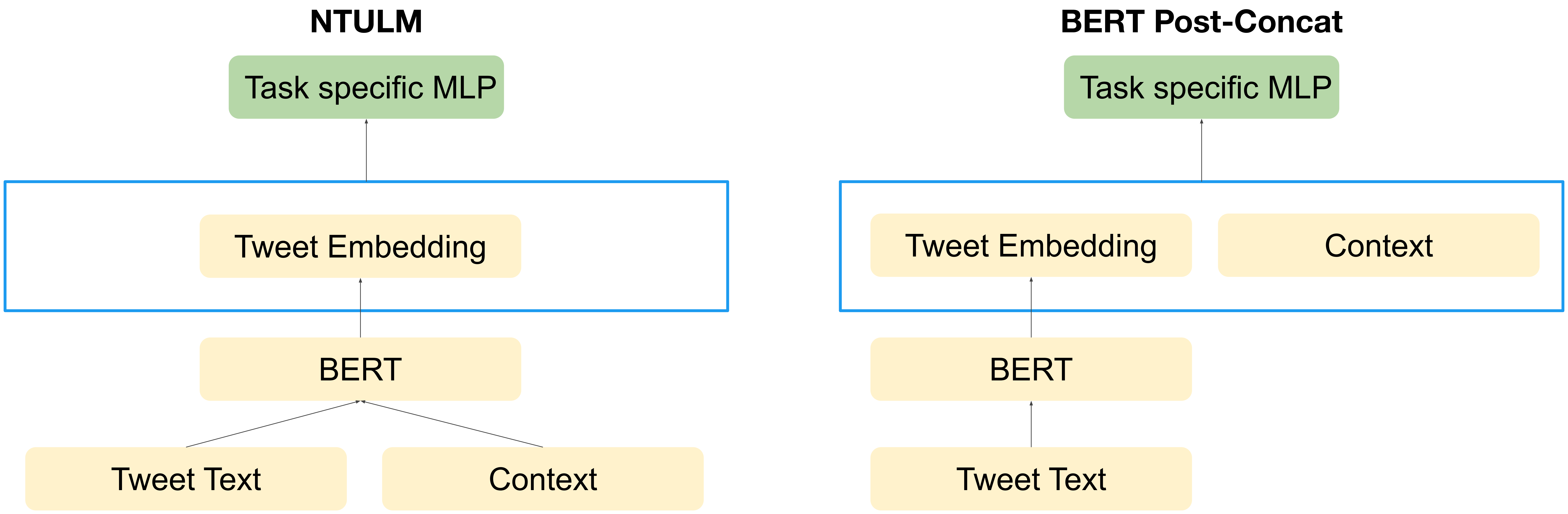


NTULM improved over BERT more when we have no OOV NTUs

Even for no NTUs, NTULM learns good text based embeddings which show small improvements.

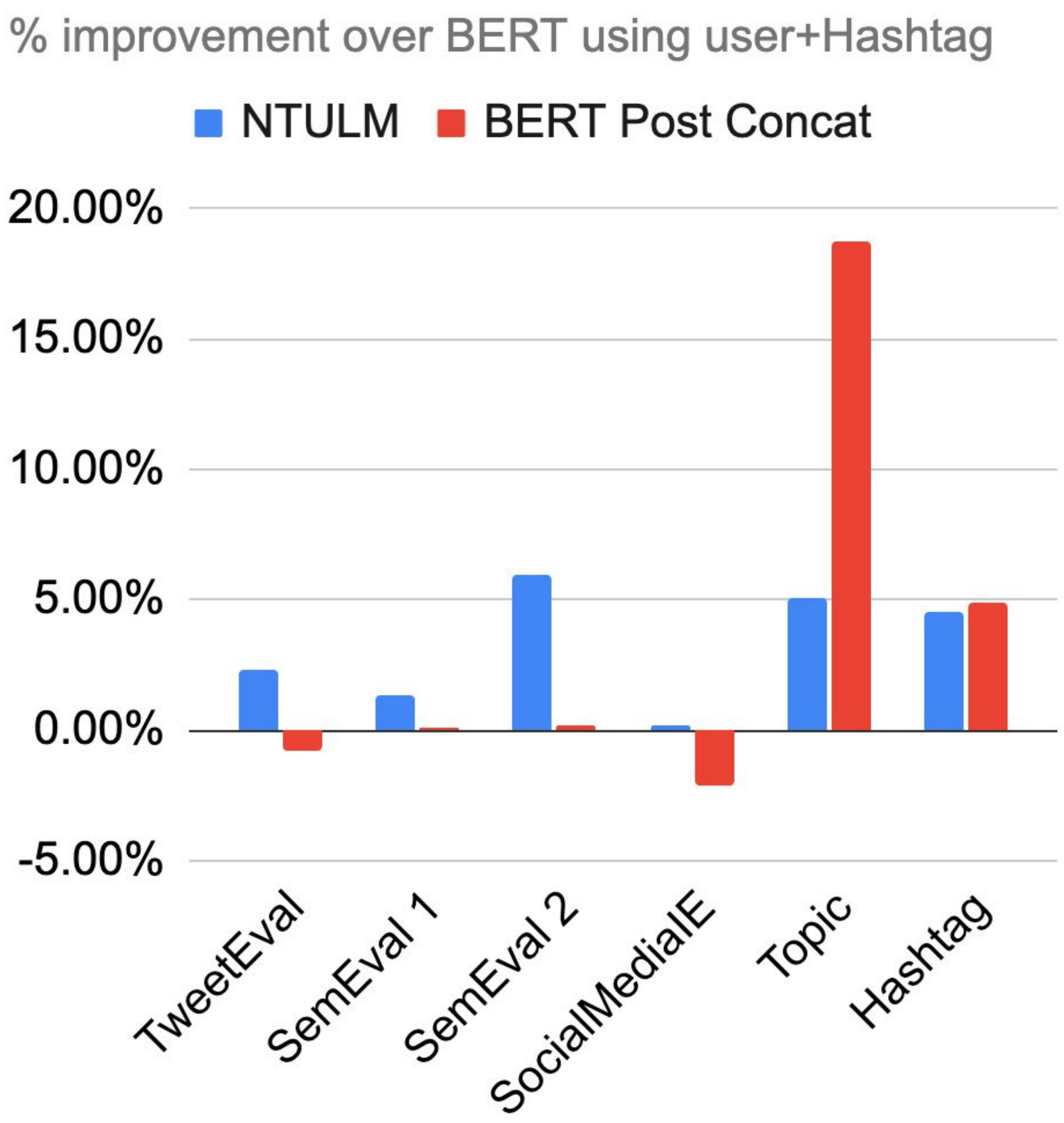
NTULM v/s BERT and Context separate

Alternative way to add context embedding: concatenate the context embedding after the BERT encoder? (named BERT Post-Concat or BERTC)



NTULM v/s BERT and Context separate

Dataset	Overall		Overlap		Non-Overlap	
	NTULM	BERTC	NTULM	BERTC	NTULM	BERTC
TweetEval	2.27%	-0.80%	2.73%	-3.33%	0.31%	0.65%
SemEval 1	1.36%	0.08%	2.59%	0.21%	0.65%	0.02%
SemEval 2	5.93%	0.22%	-0.07%	0.58%	2.62%	0.07%
SocialMediaIE	0.20%	-2.12%	-0.27%	-4.12%	1.98%	-22.22%
Hashtag	4.51%	4.87%	5.61%	7.46%	1.01%	-3.37%
Topic	5.10%	18.72%	6.92%	34.72%	0.71%	-4.17%



- **NTULM** integrates contexts embedding before attention layer, enabling the BERT encoder to automatically learn the attention of context embeddings.
- **BERTC** directly attach the context embedding after encoder, making it over-dependent on context embedding (affects the language model itself)

Recap

- NTULM shows how to integrate social context of Non Textual Units into language models
- NTULM led to significant improvements on a variety of tasks over other baselines
- Improving coverage of NTUs may further improve NTULM.

Questions

Jinning Li, Shubhanshu Mishra, Ahmed El-Kishky, Sneha Mehta, and Vivek Kulkarni. 2022. NTULM: Enriching Social Media Text Representations with Non-Textual Units. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 69–82, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Reach out on Twitter at @TheShubhanshu

References

1. Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational Sociolinguistics: A Survey. Computational Linguistics, 42(3):537–593.
2. Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media, pages 42–49.
3. Shubhanshu Mishra et al. 2018. Detecting the correlation between sentiment and user-level as well as text-level meta-data from benchmark corpora. In Proceedings of the 29th on Hypertext and Social Media, pages 2–10.
4. Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. Cognitive science, 43(1):e12701.
5. KI-BERT: Infusing Knowledge Context for Better Language and Domain Understanding, Arxiv 2021
6. K-BERT: Enabling Language Representation with Knowledge Graph, AAAI 2020
7. KEPLER: A unified model for knowledge embedding and pre-trained language representation, ACL 2021
8. Kulkarni, Vivek, Kenny Leung, and Aria Haghighi. "CTM--A Model for Large-Scale Multi-View Tweet Topic Classification." arXiv preprint arXiv:2205.01603 (2022).
9. Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofia Samaniego, Ying Xiao, and Aria Haghighi, TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation, ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (SIGKDD 2022)
10. Kulkarni, Mishra, and Haghighi. "LMSOC: an approach for socially sensitive pretraining" (2021)
11. Mishra, Shubhanshu. 2020. “Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data.” University of Illinois at Urbana-Champaign. https://shubhanshu.com/phd_thesis/.