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]

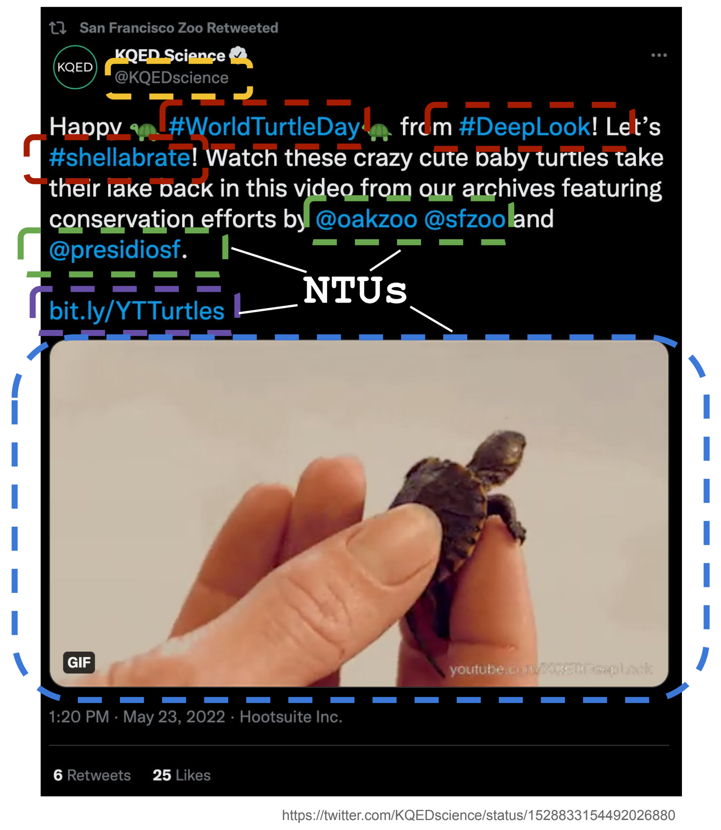
[1]University of Illinois at Urbana-Champaign, [2]Twitter, Inc., ^Equal Contribution, *Work done during internship at Twitter, Inc.

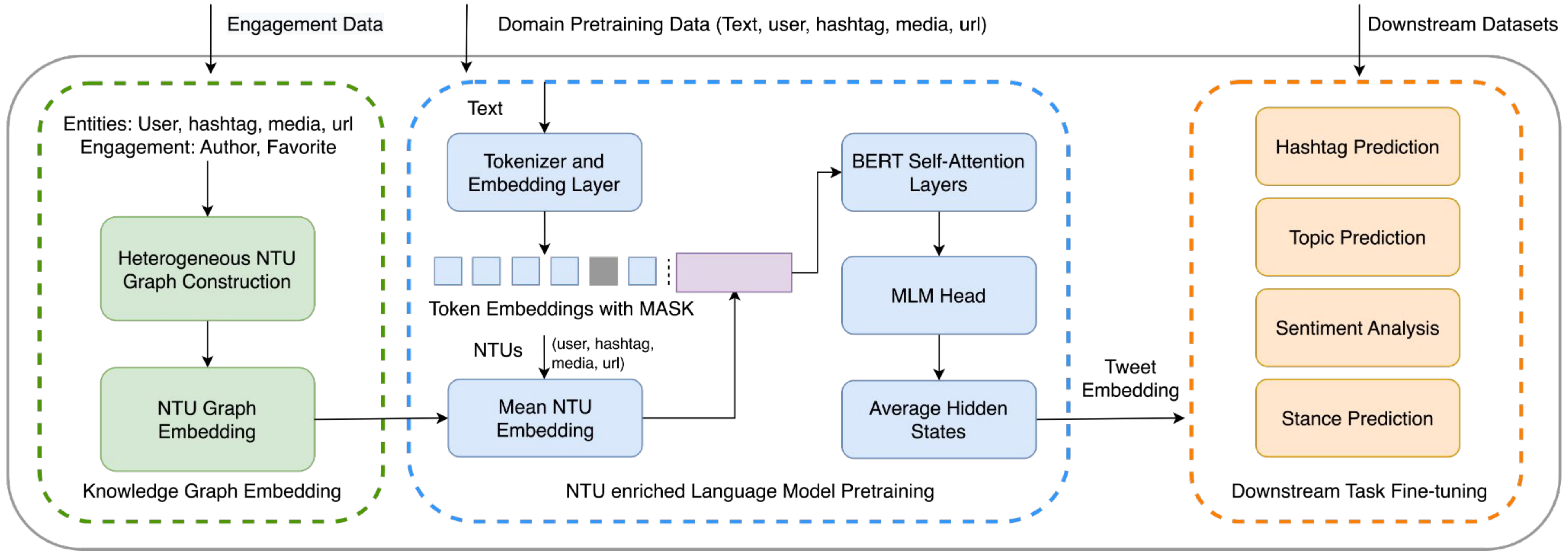
2022 The 8th Workshop on Noisy User-generated Text (W-NUT)

Non-Textual Units (NTUs)

are the social contexts which appear alongside a social media post, e.g. *Hashtag*, *URL*, *author*, *user mentions* and *media*

```
[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]
```





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)

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

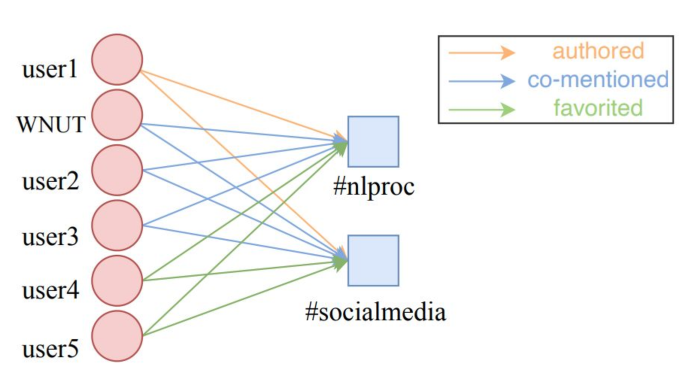


Figure 2: Graph construction with the example data in Table 1 for training NTULM user-Hashtag embeddings.

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

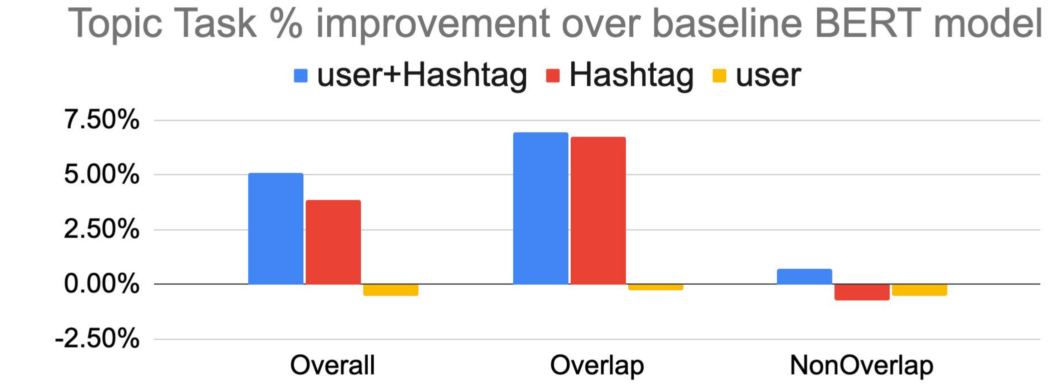
Results

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**

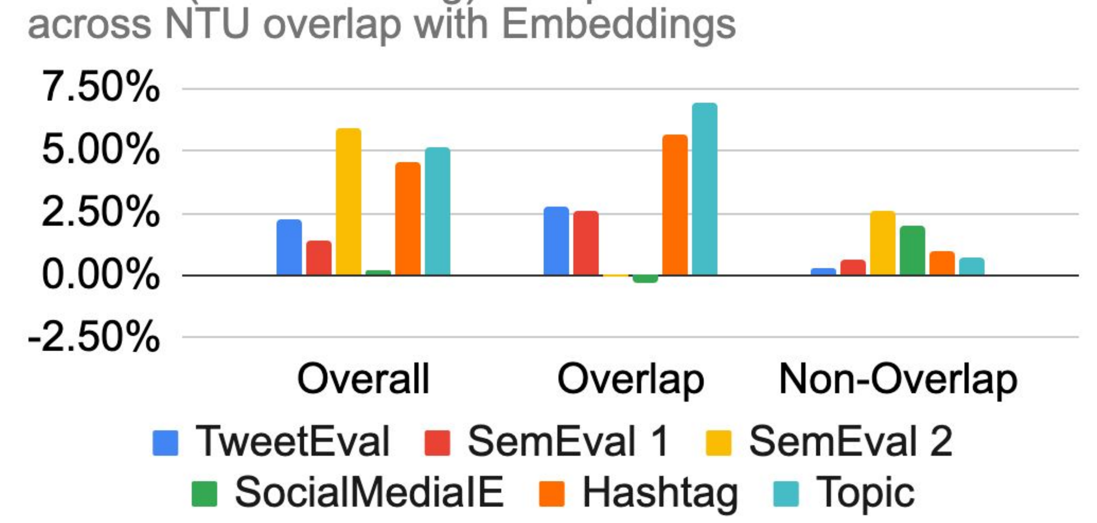
Why is NTULM effective?

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.

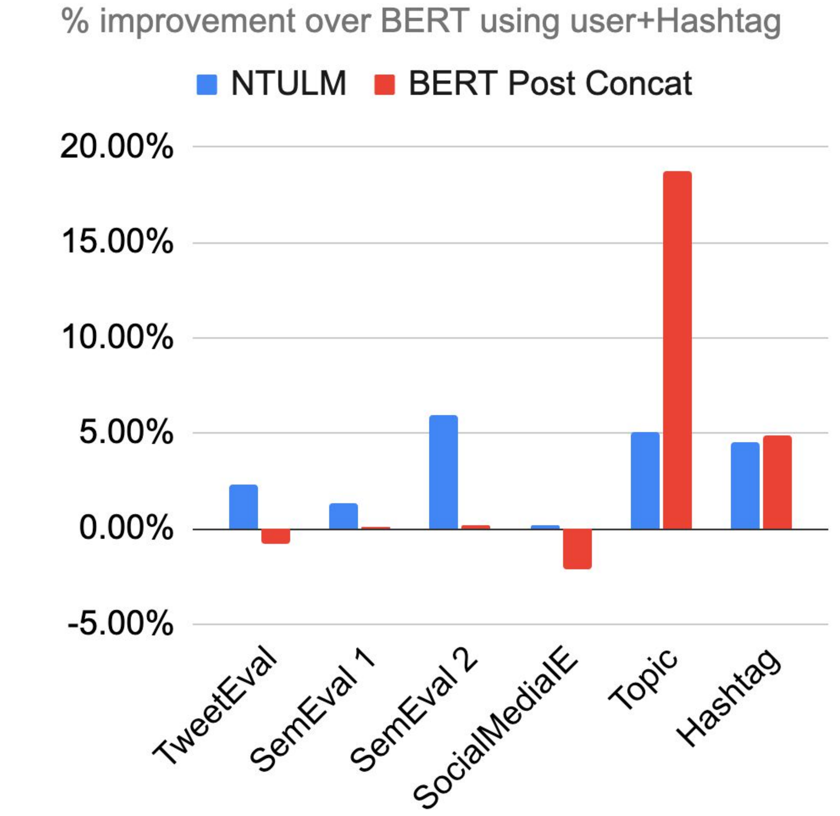


Dataset	Hashtag overlap	User overlap
Hashtag	99%	10%
SemEval	92%	21%
Social Media IE	95%	22%
Topic	99%	14%
TweetEval	98%	0%
Grand Total	95%	14%



Comparison with BERT Post Concat

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)

Conclusion

- 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.

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.