CUET_DiagNLP at #SMM4H-HeaRD 2026: Per-Axis TNM Staging from Pathology Reports and Opioid Impact Span Detection from Social Media

Shuva Dey, Priyangshu Barua, Dr. A. H. M. Ashfak Habib
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104001, u2104033}@student.cuet.ac.bd, ashfak@cuet.ac.bd

Abstract

In this paper, we describe systems for two #SMM4H-HeaRD 2026 shared tasks. Task 6 asks for per-axis TNM cancer staging from free-text TCGA pathology reports under severe label imbalance and long-document constraints. We fine-tune GatorTron-base separately on each axis using Focal loss with class weights and a pooled [CLS]–mean representation, reaching macro F1 of 0.700 (T), 0.774 (N), and 0.640 (M) on test set 2 against a baseline of 0.454, 0.591, and 0.554 respectively. Task 7 asks for span-level detection of opioid-related ClinicalImpacts and SocialImpacts in first-person Reddit posts. We combine DeBERTa-large and PubMedBERT (two seeds each) in a uniform-weight ensemble with boundary-aware loss, entity-replacement augmentation, and a first-person post filter, achieving strict F1 of 0.51 and relaxed F1 of 0.60, above both the task mean (0.46 / 0.55) and median (0.48 / 0.58).

1 Introduction

The #SMM4H-HeaRD 2026 shared tasks span biomedical NLP, clinical text mining, and social media health surveillance (Lopez-Garcia et al., 2026). This paper describes our systems for Task 6 (TNM cancer staging from pathology reports) and Task 7 (span detection of patient-reported opioid consequences in Reddit posts).

Both tasks share a core challenge: actionable health information buried in unstructured text. In Task 6, TNM staging is locked inside free-text TCGA pathology reports (Kefeli and Tatonetti, 2024); reports vary in length and terminology, and M1 cases account for under 7% of annotated samples. In Task 7, opioid consequences are rarely structured but described openly on social media (Dasgupta et al., 2018), with informal writing and sharp class imbalance between `ClinicalImpacts` and `SocialImpacts` complicating span detection. Both tasks require handling long documents, severe label imbalance, and fuzzy boundaries.

Our team (CUET_DiagNLP) used domain-adapted transformers with task-specific loss adjustments. For Task 6, we trained GatorTron-base independently on each TNM axis using Focal Loss with class weights. For Task 7, we combined DeBERTa-large (boundary-penalizing loss, entity-replacement augmentation) with PubMedBERT in a uniform-weight ensemble with a first-person filter to suppress false positives.

2 Related Work

Clinical NLP and cancer staging. Pathology reports routinely exceed 512 tokens, requiring models built for long-document encoding. GatorTron (Yang et al., 2022) is a large clinical language model pretrained on over 90 billion words of electronic health records, making it well-suited to the terminology and structure of TCGA pathology reports. Class-weighted cross-entropy and Focal Loss (Lin et al., 2017) are standard remedies for the label skew common in TCGA staging data, where M1 may represent under 7% of labeled examples (Kefeli and Tatonetti, 2024). Multimodal extensions of such pipelines, such as MedSAM2 (Wang et al., 2024), further broaden clinical AI toward integrated vision–language representations.

Social media NER for opioid impact detection. Clinical NER models transfer poorly to social media due to informal writing and ambiguous entity boundaries. PubMedBERT (Gu et al., 2021) and DeBERTa (He et al., 2021) are the dominant backbones for health-related span detection. Boundary-aware losses and entity-replacement augmentation provide complementary gains on minority types (Li et al., 2020; Dai and Adel, 2020). Task 7 evaluates under strict and relaxed F1, making boundary precision the key differentiator (SMM4H-HeaRD Organizers, 2026).

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 182–186

Stage	Class	Count	Total	Nulls
T	T1	1,484	5,853	921
	T2	1,985		
	T3	1,795		
	T4	589		
N	N0	2,829	4,826	1,948
	N1	1,241		
	N2	589		
	N3	167		
M	M0	3,650	3,916	2,858
	M1	266		

Table 1: Task 6 training label distribution per TNM axis. Nulls = unannotated rows excluded from that axis.

Split	Samples	Labels
Train	6,774	provided
Validation	2,279	withheld
Test set 1 (standard)	2,599	withheld
Test set 2 (hard)	499	withheld

Table 2: Task 6 dataset splits.

3 Data Description

3.1 Task 6: TNM Cancer Stage Prediction

Task 6 uses de-identified free-text pathology reports from TCGA (Kefeli and Tatonetti, 2024). Each report is labeled with three independent axes: T (tumor stage: T1–T4), N (lymph node involvement: N0–N3), and M (metastasis: M0–M1). Labels contain substantial missingness per axis, and M1 is severely underrepresented (266 of 3,916 labeled M samples). Class distributions and dataset splits are shown in Tables 1 and 2.

3.2 Task 7: Opioid Impact Span Detection

Task 7 provides first-person Reddit posts from opioid-related communities annotated with five BIO labels: O, B/I-ClinicalImpacts (e.g., withdrawal, overdose), and B/I-SocialImpacts (e.g., job loss, arrest). `SocialImpacts` spans 575 of 1,414 entity tokens (40.7%), the primary imbalance. Table 3 summarizes the splits.

Split	Sentences	w/ Entities	ClinImpact	SocImpact
Train	842	251	616	408
Dev	258	81	223	167
Test	578	—	—	—
Total	1,678	332	839	575

Table 3: Task 7 dataset statistics. Entity counts are token-level occurrences.

4 Methodology

4.1 Task 6: GatorTron-base for TNM Staging

Figure 1 illustrates the pipeline. Reports are preprocessed by expanding TNM abbreviations (`pT2`→*pathologic tumor T2*) and applying sentence dropout augmentation at training time.

Three independent classifiers are trained on `UFNLP/gatortron-base` (Yang et al., 2022), one per TNM axis, using a `[CLS]`–mean-pool GELU head with multi-dropout averaging. Training uses Focal Loss (γ=2, label smoothing 0.05) with per-class weights and AdamW with a cosine schedule.

At inference, the best checkpoint per axis by validation macro-F1 is selected, with M-stage threshold tuning to address M1 imbalance. Hyperparameters are in Table 4.

Parameter	Value
Max length (T, N)	512
Max length (M)	256
Learning rate	2×10^{-5}
Batch size	4
Grad. accum. steps	8
Effective batch size	32
Max epochs	10
Early-stop patience	3
Weight decay	0.01
Warmup ratio	0.06
Dropout	0.10
Seed	42

Table 4: Task 6 hyperparameters (`GatorTron-base`) identical across axes except max length (512 T/N, 256 M).

4.2 Task 7: Boundary-Aware DeBERTa + PubMedBERT Ensemble

We fine-tune DeBERTa-large (He et al., 2021) and PubMedBERT (Gu et al., 2021) (two seeds each) for token classification with three targeted modifications (Figure 1).

Boundary-aware loss. $\mathcal{L}_B$ is cross-entropy restricted to gold B- tokens, applying extra penalty on span-start misses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + (\lambda - 1)\mathcal{L}_B, \quad \lambda{=}1.5 \qquad (1)$$

SocialImpacts class weights are upscaled by 1.35 with label smoothing ϵ=0.05. Both models are trained with this loss.

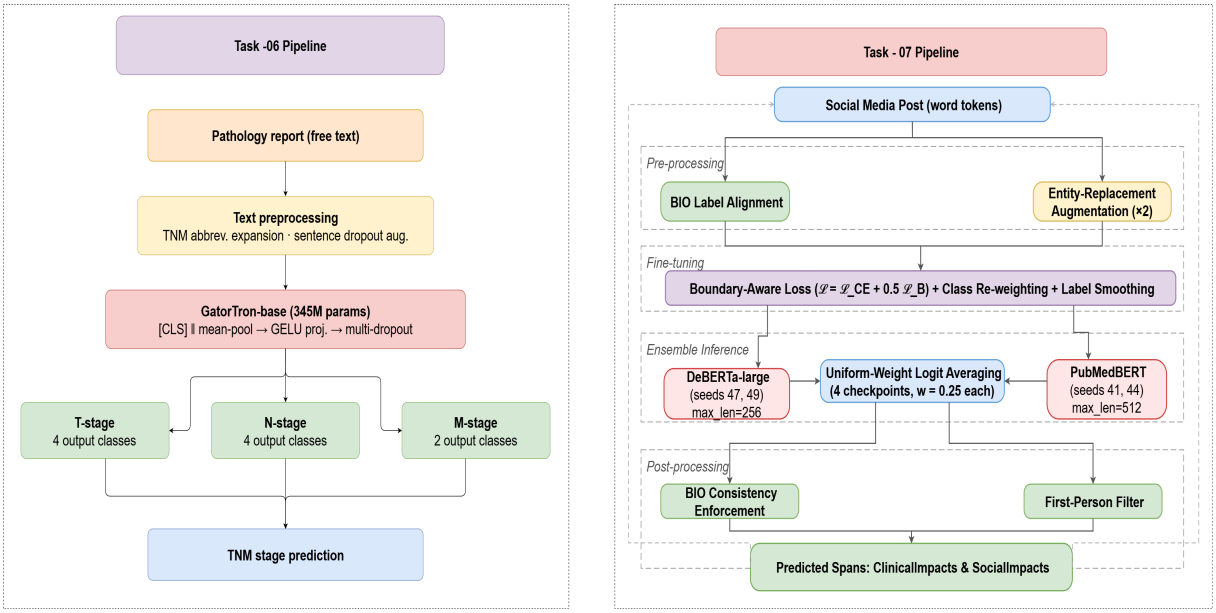

Figure 1: Pipeline overview for Task 6 (left) and Task 7 (right)

Hyperparameter	DeBERTa-large	PubMedBERT
Max sequence length	256	512
Physical batch size	4	8
Grad. accumulation steps	8	4
Effective batch size	32	32
Learning rate	2×10^{-5}	3×10^{-5}
Weight decay	0.01	0.01
Warm-up ratio	0.15	0.10
Dropout p	0.20	0.10
Max epochs	10	10
Early-stop patience	5	5
Gradient clip norm	1.0	1.0
Label smoothing ϵ	0.05	0.05
Seeds	47, 49	41, 44

Table 5: Task 7 Hyperparameter configurations for DeBERTa-large and PubMedBERT.

Entity-replacement augmentation. Labeled spans are substituted with phrases from curated synonym pools (23 clinical, 20 social), expanding training data by ∼30%.

First-person filter. Posts lacking self-referential markers (*i, i'm, my, me, myself, we, our*) receive all-O predictions at inference, suppressing false positives on third-person text.

The final ensemble averages word-level logits equally across all four checkpoints—two DeBERTa-large seeds and two PubMedBERT seeds—followed by BIO consistency enforcement in post-processing. Full hyperparameter configurations are reported in Table 5.

5 Results and Analysis

5.1 Task 6 Results

Table 6 reports GatorTron-base results across all evaluation sets. On the validation set (2,279 samples) the model reaches macro F1 of 0.774 and micro F1 of 0.874, correctly identifying 870/1034 T-stage, 743/852 N-stage, and 641/692 M-stage labels. M-stage lags behind T and N (0.680 vs. 0.836/0.807) due to severe M1 underrepresentation. On test set 1 (2599 samples) the model scores substantially higher across all axes (macro F1 = 0.954, micro F1 = 0.973), possibly reflecting cleaner report structure. On test set 2 (499 samples) GatorTron-base achieves macro F1 of 0.705 and micro F1 of 0.768, outperforming the provided baseline by 0.172 macro F1 points (0.705 vs. 0.533).

5.2 Task 7 Results

Table 7 reports validation and test results. For development, models were trained on the training set only; the final submission used the combined train+dev set.

On validation, the 5-seed ensemble reaches strict F_1 of 0.463 and relaxed F_1 of 0.598, covering 207/390 gold spans. `ClinicalImpacts` precision (0.691) and coverage (114/223) exceed `SocialImpacts` (0.679, 93/167), consistent with its higher training frequency.

On the test set, the 4-model DeBERTa+PubMedBERT ensemble achieves strict F_1 of 0.51 and relaxed F_1 of 0.60, above the task mean

Evaluation set	Model	T F1	T Correct	N F1	N Correct	M F1	M Correct	Macro F1	Micro F1
Validation set	GatorTron-base	0.836	870/1034	0.807	743/852	0.680	641/692	0.774	0.874
Test set 1	GatorTron-base	0.976	97/100	0.960	96/100	0.926	99/100	0.954	0.973
Test set 2	GatorTron-base	0.700	—	0.774	—	0.640	—	0.705	0.768
Baseline (test set 2)	—	0.454	—	0.591	—	0.554	—	0.533	0.517

Table 6: Task 6 macro-F1 and correct predictions per TNM axis.

Split	System / Entity	P	R	Strict F_1	Relaxed F_1	Coverage
Val	ClinicalImpacts (DeBERTa 5-seed)	0.691	0.511	—	—	114/223
	SocialImpacts (DeBERTa 5-seed)	0.679	0.557	—	—	93/167
	Overall (DeBERTa 5-seed)	0.685	0.531	**0.463**	**0.598**	207/390
Test	ClinicalImpacts — DeBERTa 5-seed	0.650	0.508	—	—	130/256
	SocialImpacts — DeBERTa 5-seed	0.649	0.463	—	—	50/108
	Overall — DeBERTa 5-seed	0.650	0.495	0.48	0.56	180/364
	ClinicalImpacts — DeBERTa+PubMedBERT	0.739	0.531	—	—	136/256
	SocialImpacts — DeBERTa+PubMedBERT	0.719	0.426	—	—	46/108
	Overall — DeBERTa+PubMedBERT	**0.734**	**0.500**	**0.51**	**0.60**	182/364
Test (ref.)	Task mean	—	—	0.46	0.55	—
	Task median	—	—	0.48	0.58	—

Table 7: Task 7 span detection results. Val: 5-seed DeBERTa-large ensemble. Test: DeBERTa-only vs. DeBERTa+PubMedBERT (primary).

(0.46 / 0.55) and median (0.48 / 0.58). Compared to the DeBERTa-only 5-seed (0.48 / 0.56), gains are precision-driven: `ClinicalImpacts` precision rises from 0.65 to 0.74 with coverage improving from 130 / 256 to 136 / 256. `SocialImpacts` recall drops from 0.463 to 0.426, suggesting PubMedBERT trades social-span recall for clinical precision.

6 Conclusion

We described two systems for #SMM4H-HeaRD 2026. Both start with a domain-adapted transformer and apply targeted loss adjustments, showing that task-specific loss design and data augmentation matter more than architectural complexity for both cancer staging from pathology text and social media NER.

For Task 6, Focal Loss with per-class weights partially addresses M-stage scarcity, though M1 underrepresentation remains the binding bottleneck in TNM staging from TCGA reports. For Task 7, the DeBERTa+PubMedBERT ensemble outperforms the DeBERTa-only system; false negatives on long `SocialImpacts` spans remain the primary gap from the organizer benchmark.

Future directions include span-level objectives for Task 7 and semi-supervised learning over unlabeled pathology reports for Task 6.

Limitations

For Task 6, GatorTron-base truncates reports to 512 tokens (256 for M-stage), dropping distal sections where staging evidence often appears. M1 is severely underrepresented in TCGA, and Focal Loss only partially compensates. Training on a single Kaggle T4 GPU (15 GB) prevented ensembling GatorTron with Clinical-Longformer due to session runtime constraints, which was planned but had to be dropped because of session runtime limits.

For Task 7, DeBERTa-large's 256-token window truncates long `SocialImpacts` spans, directly depressing recall; PubMedBERT's 512-token limit still misses multi-sentence spans. The synonym pools (23 clinical, 20 social) are manually built and miss rare phrasings. The first-person filter (*i, i'm, my, me, myself, we, our*) suppresses predictions on posts without explicit markers, trading recall for precision. With only four checkpoints across two architectures, the ensemble is narrow; more seeds would reduce variance.

References

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3861–3867.

Nabarun Dasgupta, Leo Beletsky, and Daniel Ciccarone. 2018. Opioid crisis: No easy fix to its social and economic determinants. *American Journal of Public Health*, 108(2):182–186.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jenna Kefeli and Nicholas Tatonetti. 2024. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. *Patterns*, 5(3):100933.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5849–5859.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, et al. 2026. Overview of the 11th social media mining for health (#smm4h) and health real-world data (HeaRD) shared tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

SMM4H-HeaRD Organizers. 2026. Task 7 data description: Extraction of social and clinical impacts of substance use from social media posts. SMM4H-HeaRD 2026 Shared Task Description. `https://www.codabench.org/competitions/13991`.

Jun Wang et al. 2024. Medsam2: Segment medical images as video via segment anything model 2. *Nature Communications*, 15(1):9003.

Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2022. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.