Team ID: 23
Team Name: Solo Project

Evaluating BERT-Based Models for Negation Detection in Biomedical Texts

UNIVERSITÉ Concordia UNIVERSITY

Abstract

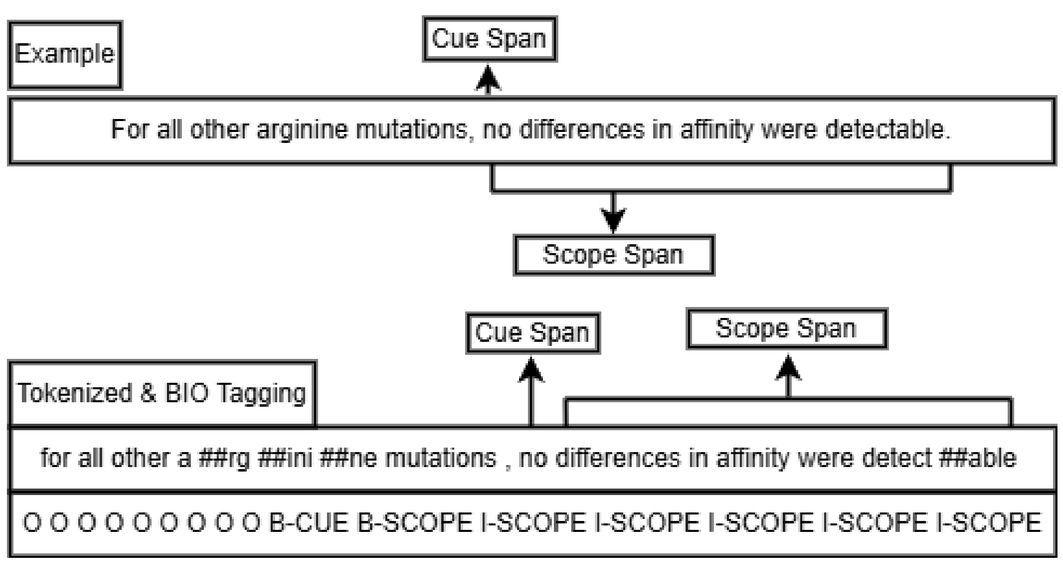
- Negation and speculation in biomedical texts affect the interpretation of clinical findings and and their downstream analysis.
- This project explores how contextual embeddings from BERT-based models can detect negation cues and scopes in sequence labeling tasks.

Project Objective

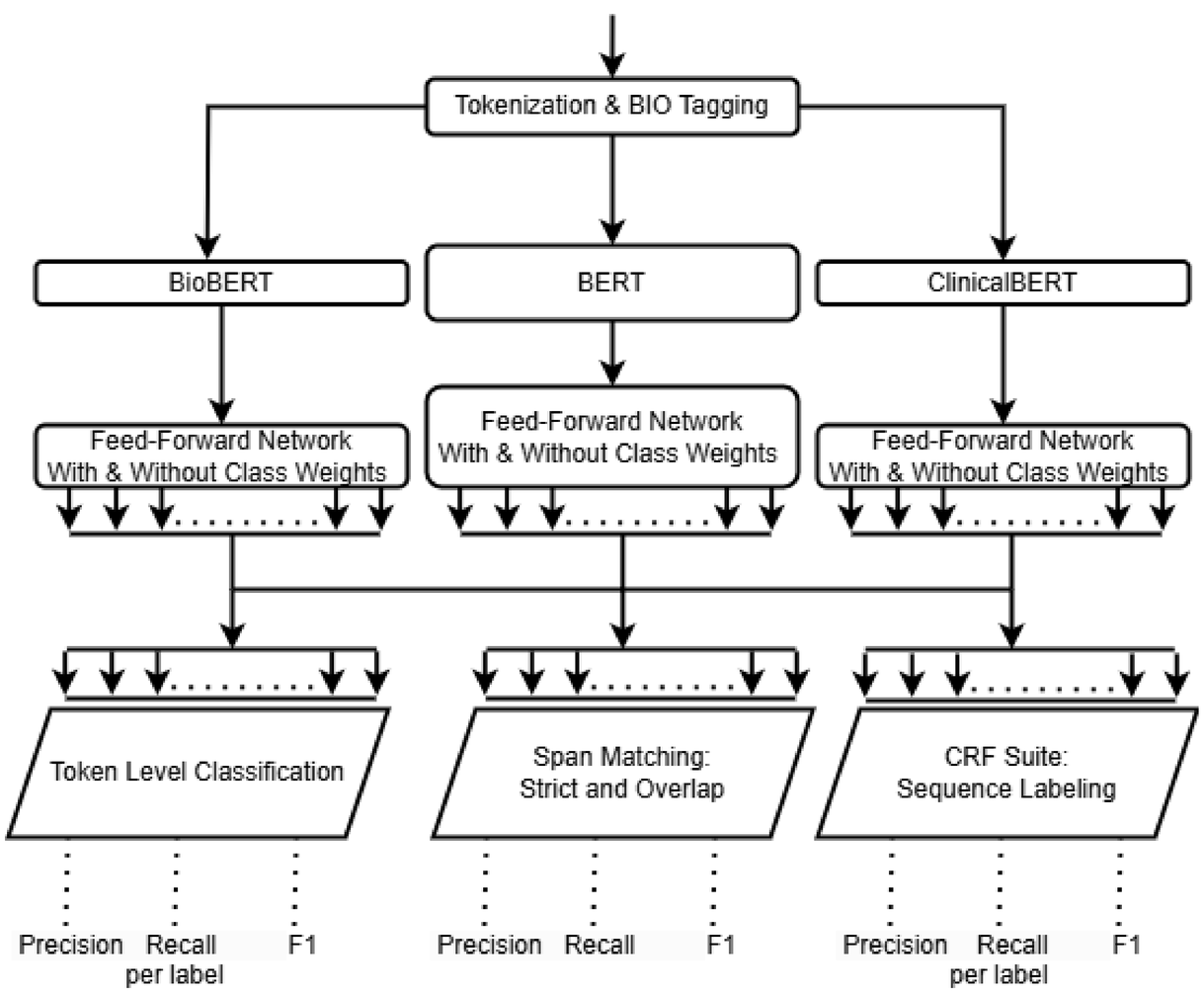
- Convert cue and scope spans in the BIOSCOPE dataset [1] into sequence-labeled BIO tags: (`B-CUE`, `I-CUE`, `B-SCOPE`, `I-SCOPE`, `O`).
- Fine-tune BERT-based models on BIO tagged data to evaluate their ability to detect negation cues and scopes.

Methodology

- Cue spans were removed from scope spans following [2].
- Fine-tuned BioBERT, BERT, and ClinicalBERT using BIO-tagged labels with a Feed Forward Network.
- Loss was calculated without and with class weights (Methods 1 and 2) to address label imbalances.
- Performance Evaluated on Token-Level Classification, Span Matching, and CRF Suite.



Example: Original sentence, tokenized text, and BIO-tagged labels.



Proposed architecture for fine-tuning and evaluation.

Results

Model	Method	Token-Level	Span Matching		CRF Suite
		Micro F1 %	Strict F1 %	Overlap F1 %	Micro F1 %
BioBERT	1	98.01	18.12	53.82	60.92
BERT	1	**98.14**	**21.03**	**57.24**	**63.07**
ClinicalBERT	1	97.85	16.77	52.10	57.52
BioBERT	2	96.34	12.65	41.59	50.86
BERT	2	**96.59**	**15.42**	**44.93**	**53.81**
ClinicalBERT	2	96.31	13.43	40.78	50.62

Performance of BERT-based models.

Sample Predictions: True vs. Predicted Labels

True Labels:

`This, recognition, does, not (B-CUE), require (B-SCOPE), any (I-SCOPE), of (I-SCOPE), the (I-SCOPE), known (I-SCOPE), roX (I-SCOPE), RNAs (I-SCOPE).`

Predicted Labels:

`This, recognition, does, not (B-CUE), require (I-SCOPE), any (I-SCOPE), of (I-SCOPE), the (I-SCOPE), known (I-SCOPE), roX (I-SCOPE), RNAs (I-SCOPE).`

Evaluation

- **Token-Level:** Measures the F1-score for each token in the sequence.
- **Span Matching:** Evaluates spans using Strict Match (exact) and Overlap Match (partial) criteria.
- **CRF Suite:** Assesses sequence-level F1-scores by comparing predicted sequences to ground truth.

Analysis

- **BERT** (Method 1) achieves the highest scores across tasks.
- Class weights improve performance on underrepresented labels but reduce overall scores.
- **BioBERT** ranks second, while **ClinicalBERT** underperforms.
- **BERT** embeddings are most effective for detecting negation cues and scopes.

Limitations and Future Work

Limitations:

- ClinicalBERT was used due to the unavailability of a pretrained NegBERT model.
- Dataset imbalance, particularly for negation-related labels, impacted model performance.

Future Work:

- Introduce a CRF layer for enhanced sequence labeling..
- Adopt [2] unified framework to integrate multiple negation datasets.

References

[1] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, "The bioscope corpus: Biomedical texts annotated for uncertainty, negation and their scopes," *BMC bioinformatics*, vol. 9, pp. 1–9, 2008. [Online]. Available: `https://doi.org/10.1186/1471-2105-9-S11-S9`.

[2] A. Yoshida, Y. Kato, and S. Matsubara, "Negation scope conversion: Towards a unified negation-annotated dataset,", Torino, Italia: ELRA and ICCL, May 2024, pp. 12 093–12 099. [Online]. Available: `https://aclanthology.org/2024.lrec-main.1057`.

[3] M. Ballesteros, V. Francisco, A. Díaz, J. Herrera, and C. Gómez-Rodríguez, "Inferring the scope of negation in biomedical documents," in *Natural Language Processing and Information Systems*, Springer, 2012. [Online]. Available: `https://doi.org/10.1007/978-3-642-28604-9_30`.

[4] B. K. Britto and A. Khandelwal, "Resolving the scope of speculation and negation using transformer-based architectures," *arXiv preprint arXiv:2001.02885*, 2020. [Online]. Available: `https://doi.org/10.48550/arXiv.2001.02885`.