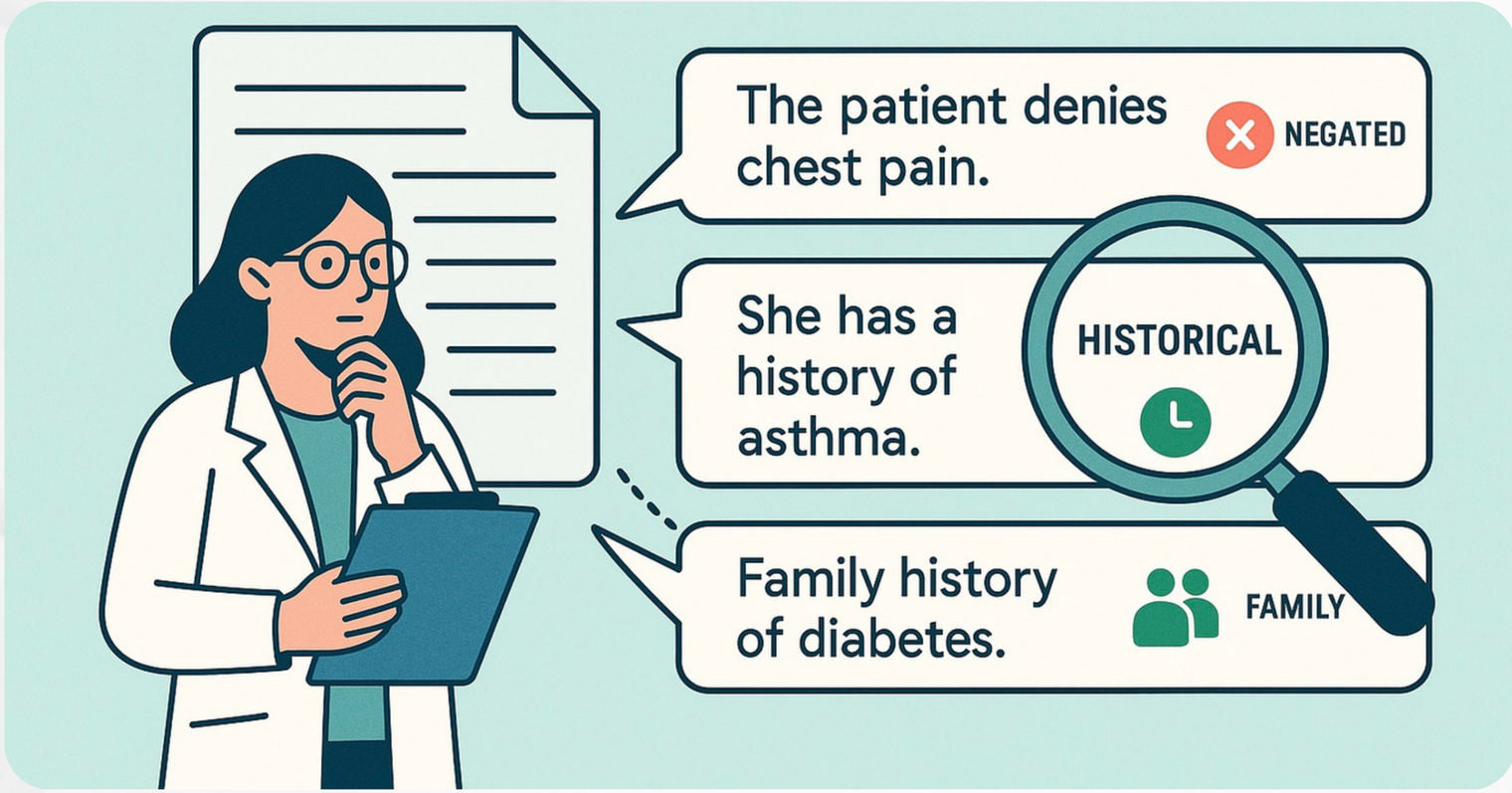
Beyond Negation Detection : Comprehensive Assertion Detection Models for Clinical NLP

Veysel Kocaman
Yigit Gul
Hasham Ui Haq
Cabir Celik
Mehmet Butgul
M. Aytug Kaya
David Talby

John Snow Labs Inc., Delaware, USA



Dataset Description

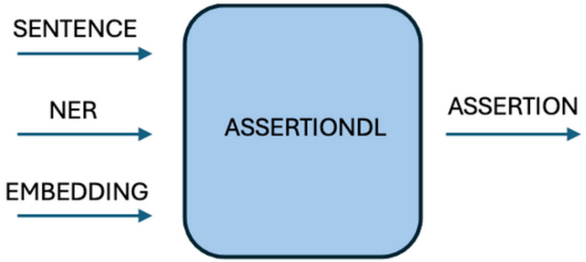
Text	Label	Description	Size
Overnight, the patient became **hypoxic** , dropping to the 80 's.	*present*	Confirms the presence of a medical condition.	8622
He gets **short of breath** with one flight of stairs.	*conditional*	Represents conditions that might occur under specific circumstances or conditions.	148
Small stroke, nearly recovered, likely **embolic from carotid artery** .	*possible*	Suggests uncertainty or potential presence of a condition.	652
There was no evidence of **diarrhea** during medical Lawrence Memorial Hospital stay.	*absent*	Indicates the negation or nonexistence of a medical condition.	2594
Mother suffer **MI** in her 50 's, died at age 59.	*associated with someone else (awse)*	Refers to medical conditions related to individuals other than the patient, such as family members.	131
Hydrocodone 5 mg with Tylenol , one to two tablets every four hours p.r.n. **pain** .	*hypothetical*	Denotes speculative or conjectural conditions that are not currently present.	445

The evaluation and benchmarking in this study are conducted exclusively on the official i2b2 dataset, which represents a comprehensive resource for assessing assertion detection frameworks in real-world clinical scenarios.

Methodology

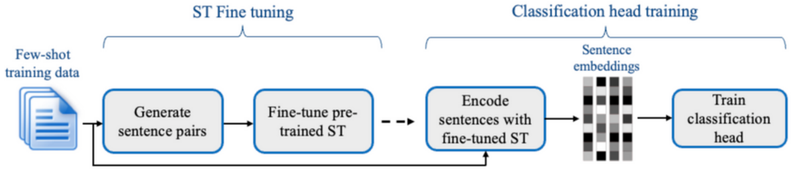
Assertion DL

AssertionDL is a Bi-LSTM-based classification model designed for assertion detection, built on a modified version of a previous architecture.



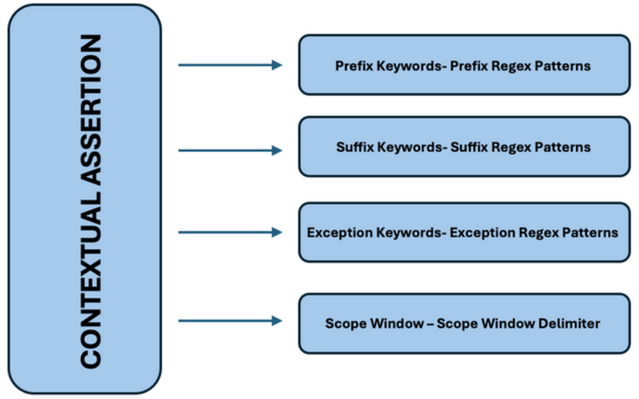
FewShot Assertion

FewShotAssertion is a transformer-based model built on a modified SetFit framework, leveraging sentence-transformer embeddings and contrastive learning for few-shot assertion detection.



Contextual Assertion

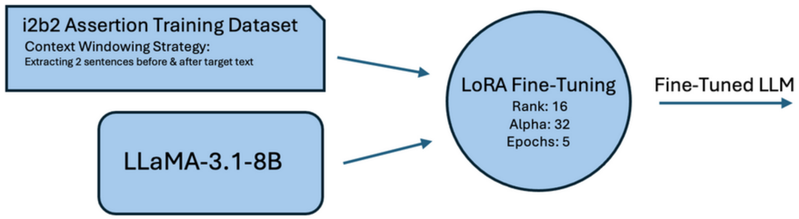
The Contextual Assertion module extends rule-based assertion detection by leveraging user-defined rules and contextual patterns



Methodology

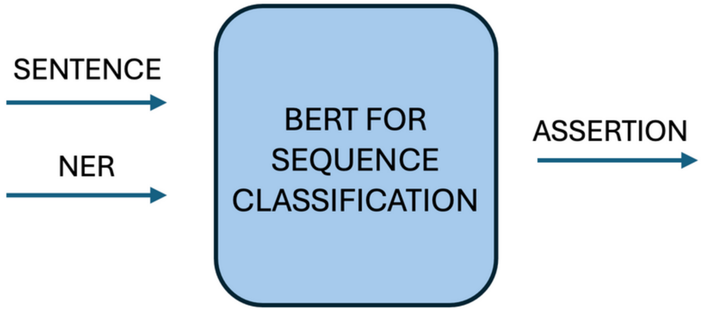
Fined-Tuned LLM

We fine-tuned LLama-3.1-8B on the i2b2 assertion training dataset using LoRA without quantization, ensuring parameter efficiency while preserving pre-trained knowledge.



Bert For Sequence Classifier

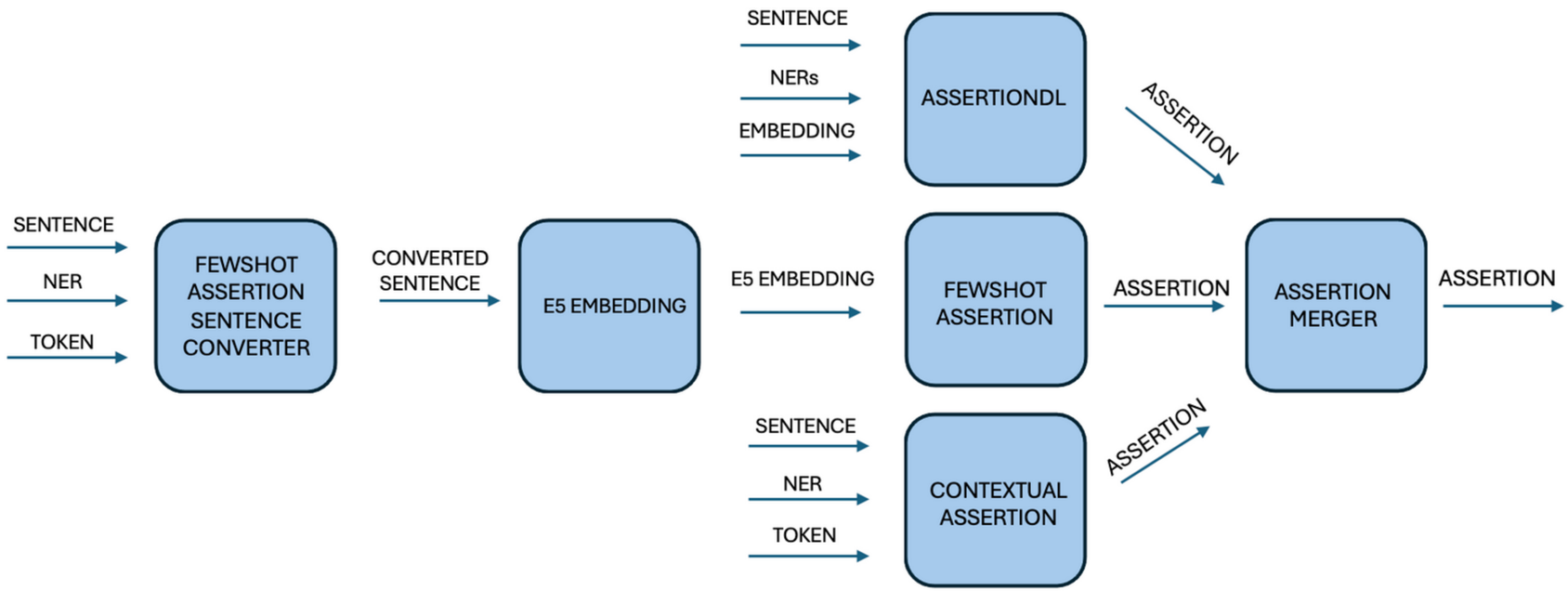
We implemented a transformer-based approach using BioBERT, a biomedical fine-tuned BERT model.



Methodology

Combined Pipeline

- **The pipeline includes AssertionDL, FewShotAssertion, and Contextual Assertion models.**
- **A majority voting mechanism is applied to resolve conflicts in predictions across models.**



Results

Model	present	absent	possible	hypothetical	conditional	awse*	weighted avg
*Combined Pipeline***	0.963	0.951	0.755	0.875	0.511	0.922	0.941
AssertionDL	0.941	0.898	0.672	0.761	**0.599**	0.886	0.907
FewShotAssertion	0.955	0.942	0.748	0.872	0.293	0.809	0.929
ContextualAssertion	-	0.929	0.708	-	-	0.835	0.883
Fine Tuned LLM	**0.976**	**0.975**	0.759	0.911	-	**0.943**	**0.962**
BFSC (BioBert)	0.975	0.972	**0.787**	0.918	0.590	0.913	0.957
GPT-4o	0.937	0.891	0.692	0.677	-	0.805	0.901
Azure Ai Text Analytics	-	0.761	0.583	0.763	0.569	0.800	0.727
AWS Med Comprehend	0.882	0.788	0.659	0.617	-	0.737	0.839
NegEx	-	0.897	-	-	-	-	0.897
BFSC latest best [11]	0.979	0.972	0.786	-	-	-	0.952
Prompt-based Bert [29]	0.971	0.968	0.763	**0.921**	0.485	0.875	0.951

Comparison of assertion models across various categories. Best performing model for each category is represented with bold characters. The models in the first section of this table are developed by JSL. In LLM and GPT-4o experiments, hypothetical and conditional labels are merged/treated as a single label.

Conclusion

- **The study evaluates JSL's state-of-the-art assertion detection models, from lightweight DL models to fine-tuned LLMs.**
- **The fine-tuned LLM achieves the highest accuracy (96.2%), outperforming GPT-4o (90.1%) and commercial APIs, especially in Present, Absent, and Hypothetical assertions.**
- **However, the LLM is extremely costly, running 100× slower on a CPU and being thousands of times more expensive for just 1-2% better accuracy.**
- **Assertion DL, FewShot Assertion, and Bert For Sequence Classifier models offer efficient, competitive alternatives, with the Combined Pipeline (94.1%) outperforming all commercial solutions.**
- **Integrated with Spark NLP, these smaller, domain-specific models surpass GPT-4o, Azure AI, and AWS Medical Comprehend, providing scalable, cost-effective clinical NLP solutions.**

Thank You