2025 ECIR Text2Story

BEYOND NEGATION DETECTION: COMPREHENSIVE ASSERTION DETECTION MODELS FOR CLINICAL NLP

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

John Snow Labs Inc., Delaware, USA

3

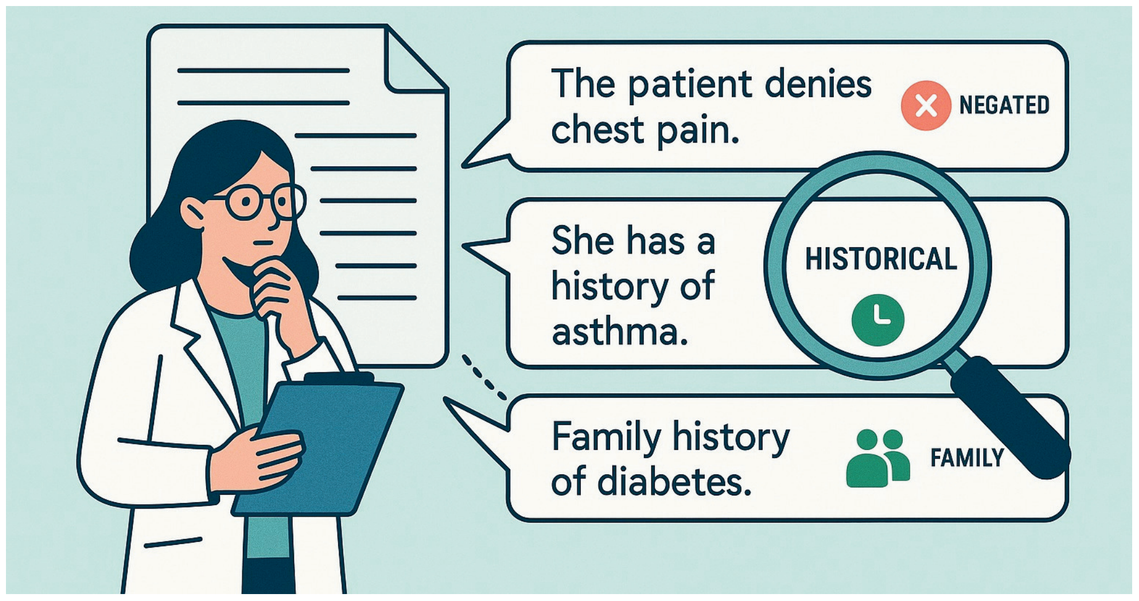
John Snow LABS

Abstract

Assertion status detection is crucial in clinical NLP but often overlooked, leading to underperformance in commercial solutions like AWS Medical Comprehend, Azure AI Text Analytics, and GPT-4o due to limited domain adaptation. To address this, we developed state-of-the-art assertion detection models, including fine-tuned LLMs, transformer-based classifiers, few-shot classifiers, and deep learning (DL) approaches. Our fine-tuned LLM achieves the highest accuracy (0.962), outperforming GPT-4o (0.901) and commercial APIs, excelling in Present (+4.2%), Absent (+8.4%), and Hypothetical (+23.4%) assertions. Our DL models surpass commercial solutions in Conditional (+5.3%) and Associated with Someone Else (+10.1%), while the few-shot classifier (0.929) offers a lightweight alternative for resource-limited settings. Integrated with Spark NLP, our models outperform black-box commercial solutions, enabling scalable inference and seamless integration with medical NLP tasks.

Data 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

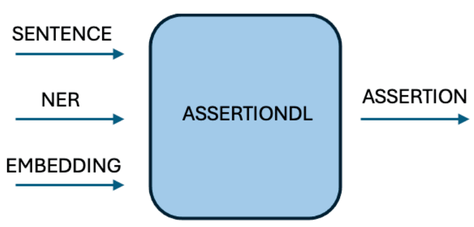


The study evaluates assertion detection models using the official 2010 i2b2 test set, a widely recognized benchmark in clinical NLP. The dataset covers six assertion categories: Absent, Associated with someone else, Conditional, Hypothetical, Possible, and Present.

Methodology

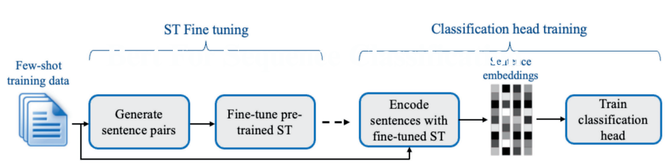
Assertion DL

- **AssertionDL is a Bi-LSTM-based classification model designed for assertion detection, built on a modified version of a previous architecture.**
- **It processes entities along with a context string, which is tokenized and embedded before being passed to the model.**



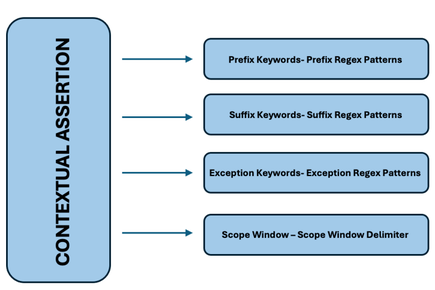
FewShot Assertion

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



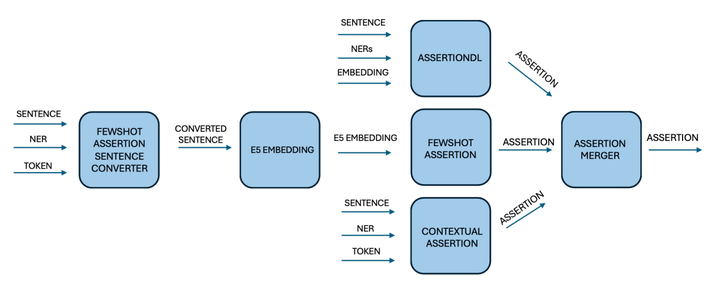
Contextual Assertion

- **The ContextualAssertion module extends rule-based assertion detection by leveraging user-defined rules and contextual patterns**
- **Inspired by ConText and NegEx.**



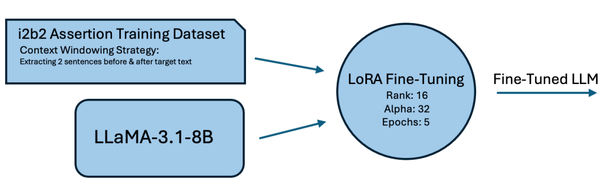
Combined Pipeline

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



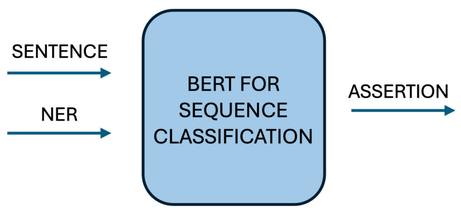
Fined-tuned LLM

- **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.**
- **Instead of processing full documents, we extracted two sentences before and after the target to improve model focus and efficiency.**



Bert For Sequence Classifier

- **We implemented a transformer-based approach using BioBERT, a biomedical fine-tuned BERT model.**
- **The model takes the target entity and its surrounding context as input.**
- **The model predicts the assertion status of the target entity.**

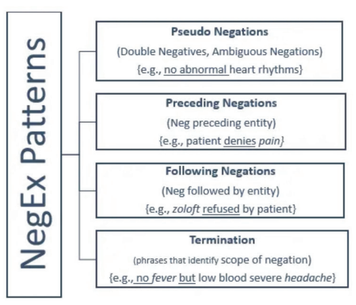


AWS \ Azure AI \ GPT-4o

- **We used GPT-4o to benchmark assertion detection for medical conditions on the i2b2 dataset.**
- **We also evaluated AWS Comprehend Medical and Azure AI Text Analytics, cloud-based NLP services designed to extract and analyze medical information from unstructured text.**
- **Due to data-sharing restrictions, we obfuscated PHI and medical terms using Healthcare NLP tools from John Snow Labs.**
- **A carefully crafted prompt guided GPT-4o in classifying assertion statuses for specified entities.**

Negex

- **NegEx is a rule-based algorithm designed to identify negation in clinical text, particularly to determine whether a medical concept is absent or not.**



Results

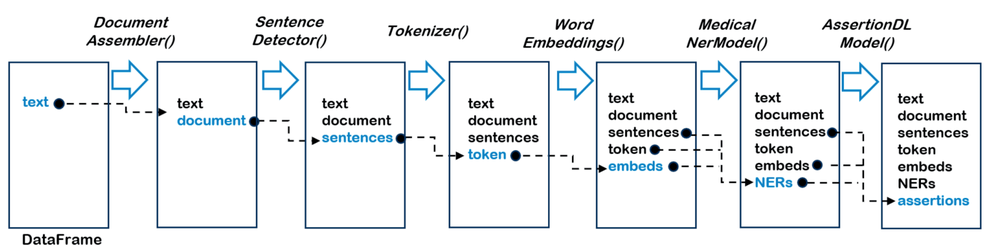
Comparison of assertion models across various categories. Best performing model for each category is represented with bold characters. The models in the first section of this table are developed by JSL. In LLM and GPT-4o experiments, hypothetical and conditional labels are merged/treated as a single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

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

***awse: associated with someone else. **Combined pipeline elements denoted in italics.**

Methods	Parameter Size	CPU(seconds)	GPU(seconds)
Fine-Tuned LLM	8 Billion	N/A	294
Combined Pipeline	N/A	12	4
AssertionDL	11 Million	3	2
FewShotAssertion	109 Million	5	2
ContextualAssertion	N/A	2	1
BFSC	110 Million	5	4
Negex	N/A	1	1

Mean latency per 100 rows, measured in seconds for various assertion methods. Experiments were run on Google Colab servers, with CPU tasks performed on a CPU instance (8vCPU @ 2.2 GHz, 50.99 GB RAM) and GPU tasks executed on an NVIDIA A100 GPU (40 GB HBM2).



The flow diagram of a Spark NLP pipeline. When we fit() on the pipeline with a Spark data frame, its text column is fed into the DocumentAssembler() transformer and a new column document is created as an initial entry point to Spark NLP for any Spark data frame. Then, its document column is fed into the SentenceDetector(), Tokenizer() and WordEmbeddings(). Now data is ready to be fed into NER models and then to the assertion model.

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.**