

**Fundamentals of Natural Language Processing (CS3233)  
Project Report**

# **Hypothaize - An Intelligent Agent For Research Paper Hypothesis Generation**

Submitted by

Harshita S – 1RVU22CSE064

Vasanth J – 1RVU22CSE186

G R Ramya – 1RVU22CSE058

Varnita S R – 1RVU22CSE184

School of Computer Science

RV University

Submitted to

Prof. Muralidhar Billa

Associate Professor

School of Computer Science

RV University

Submission date [22-04-2025]

## **Fundamentals of Natural Language Processing (CS3233)**

### **Project Report**

### **Abstract**

The advancement of Natural Language Processing (NLP) techniques has significantly improved the capacity of AI systems to process, interpret, and generate human-like text. In this project, we developed *Hypothesize*, an AI-driven research assistant designed to autonomously generate novel scientific hypotheses from research paper abstracts. The system integrates both modern deep learning-based NLP techniques using Transformer models (T5) and traditional rule-based NLP++ approaches involving keyword matching, decision trees, and handcrafted pattern substitution.

The tool fetches research papers via the arXiv API, preprocesses the abstracts, and applies multiple hypothesis generation techniques to produce potential research ideas. These hypotheses are then evaluated on various metrics, including novelty, relevance, readability, and specificity. A comparative analysis was conducted to measure the effectiveness of Transformer-based NLP models against rule-based NLP++ methods. The results highlight the strengths and weaknesses of each approach, emphasizing the importance of integrating deep learning's creativity with the interpretability of traditional NLP. The project demonstrates how a hybrid system can support researchers in generating insightful, testable hypotheses for future investigation.

### **Objectives**

1. To design and develop an AI-driven research assistant capable of generating scientific hypotheses from research paper abstracts.
2. To implement and integrate both **modern NLP (deep learning-based)** and **traditional NLP++ (rule-based)** approaches within a unified platform.
3. To evaluate the generated hypotheses using metrics such as **BERTScore**, **semantic similarity**, **novelty**, **readability**, and **specificity**.
4. To perform a **comparative study between NLP and NLP++ techniques**, highlighting their respective advantages, limitations, and application scenarios.
5. To provide an interactive, user-friendly web application interface via **Streamlit** for exploring, generating, and evaluating hypotheses.

## **Fundamentals of Natural Language Processing (CS3233)**

### **Project Report**

## **Methodology**

This section details the end-to-end workflow and the techniques employed in the project.

### **3.1 Data Collection**

- The abstracts of scientific papers were sourced using the arXiv API based on user-defined categories and optional search queries. The API returns paper metadata including the title, abstract, authors, published date, and URL.

### **3.2 Text Preprocessing**

- Preprocessing was essential to standardize text data before feeding it into hypothesis generation models. The preprocessing steps included:
  - Lowercasing text.
  - Removing numbers, punctuation, and special characters.
  - Tokenizing text into words and sentences using NLTK and SpaCy.
  - Removing stopwords.
  - Lemmatizing tokens for their base forms.

Named Entity Recognition (NER) via SpaCy for extracting important entities.

### **3.3 Hypothesis Generation Methods**

- **Modern NLP Approach (Transformer-Based Deep Learning)**
  - Employed the T5-base transformer model fine-tuned for text generation.
  - Used custom prompts to instruct the model to generate hypotheses from abstracts.
  - Controlled generation quality using parameters like temperature, top\_k, top\_p, and max\_length.
  - Generated hypotheses were semantically rich, fluent, and varied.
  - Traditional NLP++ Approaches
- **Rule-Based Pattern Substitution:**
  - Predefined hypothesis sentence templates.
  - Populated placeholders with extracted keywords, entities, and domain-specific terms.
- **Decision Tree Heuristic:**
  - Extracted abstract features (sentence length, domain terms, frequencies).
  - Decision tree logic mapped feature values to hypothesis types.
  - Selected a hypothesis template accordingly.
- **Keyword Matching:**
  - Scored sentences based on the presence of hypothesis-related keywords.
  - Selected high-scoring sentences and transformed them into hypotheses using prefix templates.

## Fundamentals of Natural Language Processing (CS3233) Project Report

### 3.4 Evaluation Metrics

- Each generated hypothesis was evaluated using:
- BERTScore: Evaluates semantic similarity between the abstract and generated hypothesis.
- Semantic Similarity: Cosine similarity of sentence embeddings from SentenceTransformer.
- Novelty Score: Inverse similarity of hypothesis to other abstracts in the dataset.
- Readability: Approximation of Flesch-Kincaid grade level using sentence and word metrics.
- Specificity: Count of precise terms, numbers, comparisons, and scientific terminology.

### 3.5 Visualization & Comparative Analysis

- Visualization plots included:
- Radar charts for multi-metric hypothesis evaluation.
- Bar charts for comparing different generation methods.
- Processing time charts to compare efficiency.

### Tools & Libraries Used:

Tool/Library	Purpose
Python 3.x	Core programming language
Streamlit	Interactive web app interface
arXiv API	Fetching research papers
NLTK, SpaCy	Text preprocessing and NLP tasks
Transformers (T5)	Deep learning model for text generation
Sentence-Transformers	Sentence embedding and similarity scoring
BERTScore	Semantic similarity metric
Scikit-learn	Decision Tree Classifier, TF-IDF vectorization
Matplotlib, Seaborn	Data visualization and plotting

## Fundamentals of Natural Language Processing (CS3233) Project Report

### Comparative Analysis:

#### Comparison Between NLP vs NLP++

#### Modern NLP (Transformer) vs Traditional NLP++ Techniques

Feature	NLP (T5 Transformer)	NLP++ (Rule-based, Decision Tree, Keyword)
<b>Flexibility</b>	Learns complex patterns autonomously	Manually defined rules and templates
<b>Creativity</b>	High — generates diverse, novel hypotheses	Limited by rule/template variety
<b>Interpretability</b>	Low — model decisions are opaque	High — rule decisions are transparent
<b>Processing Time</b>	Higher (GPU beneficial)	Faster on CPU
<b>Novelty</b>	Typically higher due to deep context understanding	Moderate — reliant on existing keywords/patterns
<b>Readability</b>	Natural, fluent hypotheses	May vary depending on rule structure
<b>Specificity</b>	High — integrates domain-specific language	Varies based on keyword and entity extraction
<b>Use Cases</b>	General-purpose, creative hypothesis generation	Domain-specific, explainable, resource-efficient

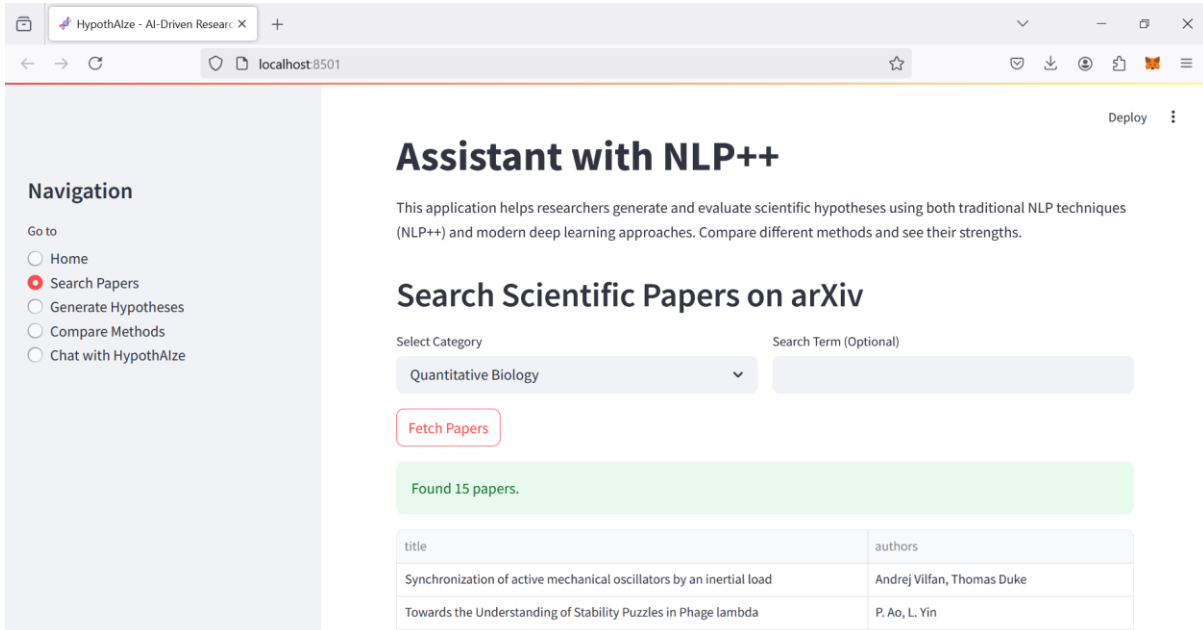
### Results

- Transformer-based NLP produced hypotheses with higher **novelty** and **readability** scores.
- NLP++ methods demonstrated faster processing times and higher **interpretability**.
- Decision Tree and Rule-Based NLP++ performed well for well-defined domains like biology and medicine.

## Fundamentals of Natural Language Processing (CS3233) Project Report

- Specificity was highest in hypotheses generated via Keyword Matching when domain terms were dense.
- Processing time: Transformer > Decision Tree > Rule-Based  $\approx$  Keyword Matching

### Output:



**Assistant with NLP++**

This application helps researchers generate and evaluate scientific hypotheses using both traditional NLP techniques (NLP++) and modern deep learning approaches. Compare different methods and see their strengths.

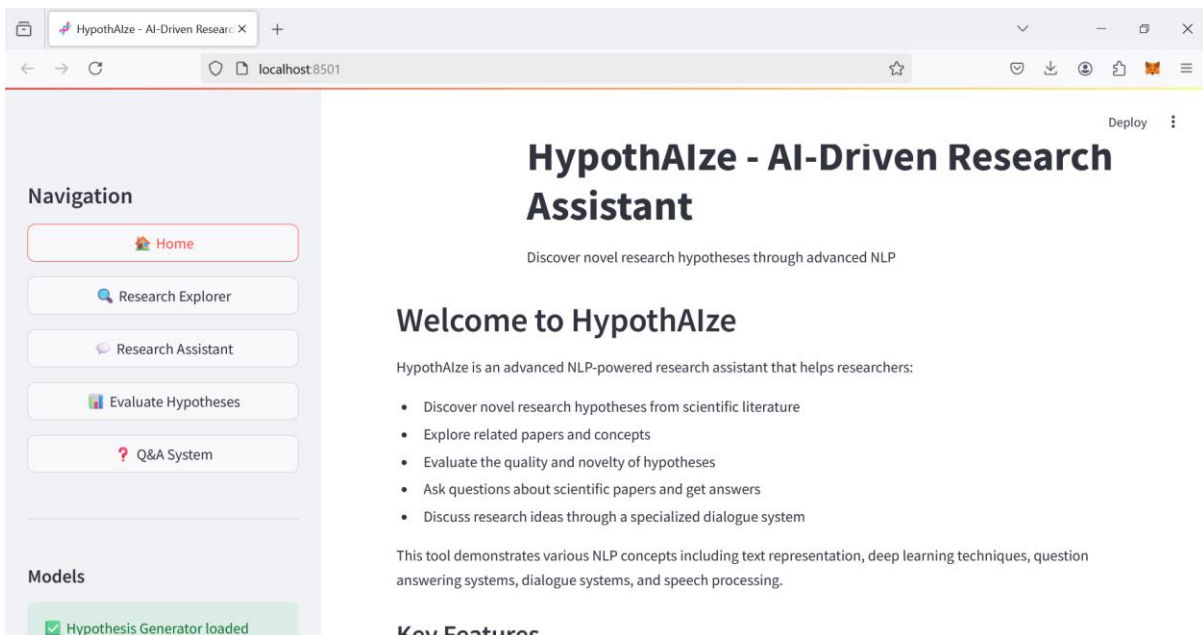
**Search Scientific Papers on arXiv**

Select Category: Quantitative Biology      Search Term (Optional):

**Fetch Papers**

Found 15 papers.

title	authors
Synchronization of active mechanical oscillators by an inertial load	Andrej Vilfan, Thomas Duke
Towards the Understanding of Stability Puzzles in Phage lambda	P. Ao, L. Yin



**HypothAlze - AI-Driven Research Assistant**

Discover novel research hypotheses through advanced NLP

**Welcome to HypothAlze**

HypothAlze is an advanced NLP-powered research assistant that helps researchers:

- Discover novel research hypotheses from scientific literature
- Explore related papers and concepts
- Evaluate the quality and novelty of hypotheses
- Ask questions about scientific papers and get answers
- Discuss research ideas through a specialized dialogue system

This tool demonstrates various NLP concepts including text representation, deep learning techniques, question answering systems, dialogue systems, and speech processing.

**Key Features**

**Navigation:**

- Home
- Research Explorer
- Research Assistant
- Evaluate Hypotheses
- Q&A System

**Models:**

- ✓ Hypothesis Generator loaded