****

**Multilingual Translator And**

**Semantic Retrieval System for**

**Sanskrit Knowledge**

**Submitted by:**

Jeevitha M

Abirami S

Dheiva Priya V

**Submitted to:**

Dr.B.Shanthi

Dean in SRC

# 

# Multilingual Translator And Semantic Retrieval

System for Sanskrit Knowledge

# 

# Objective

To develop a web-based tool that enables:

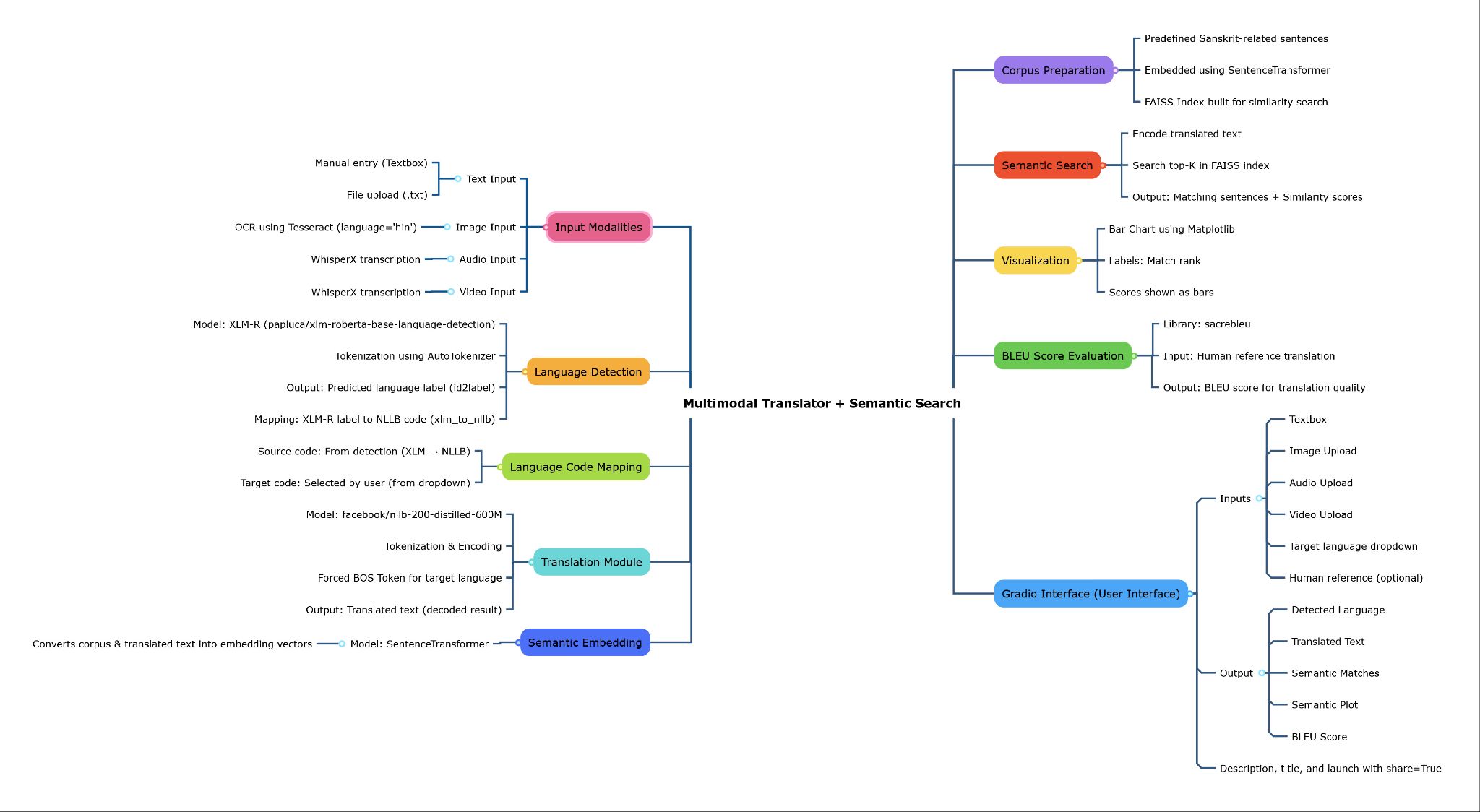
* **Automatic language detection**
* **Multilingual translation** using NLLB
* **Semantic similarity search** over curated Sanskrit-inspired corpus
* Optional **BLEU evaluation**
* Exportable **summary report**

The tool is designed for educational, linguistic, and cultural research, offering interpretive insights on translated texts via contextual matching. This project presents a multilingual translation and semantic retrieval system that goes beyond conventional translation tools. It combines automatic language detection, machine translation using Meta's NLLB-200 model, semantic sentence embedding, FAISS-based similarity search, and BLEU score evaluation. The system is capable of translating any supported language into a target language and then contextualizing it with semantically similar concepts from a curated corpus of Indic knowledge, including Sanskrit and Ayurveda-based texts. Built using Hugging Face Transformers, Sentence Transformers, and Gradio, this tool bridges linguistic and cultural gaps in natural language processing. A user-friendly web interface displays the detected language, translated output, semantic matches, similarity scores, and optional BLEU evaluation. The results are compiled into a downloadable report, offering both functionality and insight. This project provides a unique combination of translation and contextualization, adding depth to multilingual NLP.

# Introduction

The rapid growth of multilingual content across the web and academic disciplines has increased the need for effective machine translation. However, most existing tools focus purely on converting text from one language to another without considering cultural context or semantic nuance. This creates a gap, especially in domains like traditional knowledge systems, philosophy, and healthcare, where accurate meaning depends on deeper understanding.

Multilingual NLP enables systems to work across languages, yet the challenge of preserving contextual richness and indigenous semantics remains. Tools like Google Translate or DeepL may produce syntactically correct translations but often fail to relate the translation to domain-specific knowledge.



This project addresses that issue by integrating semantic search into the translation pipeline. The main objectives are:

* To detect the input language and translate text into a target language using NLLB-200
* To embed the translated text and compare it with a static corpus of culturally rich content
* To provide BLEU score-based evaluation if a human reference is available
* To deliver an interactive and downloadable output for both educational and research use

# 

# Related Work

Traditional translation tools such as Google Translate and Microsoft Translator focus on linguistic conversion and fluency. DeepL, although praised for its quality, lacks support for many low-resource or culturally complex languages. Meta's NLLB (No Language Left Behind) initiative improves multilingual coverage by supporting 200+ languages, but it remains a translation tool alone.

Moreover, tools rarely incorporate semantic comparison or domain-specific insight. There is minimal integration of semantic search engines with translation pipelines, especially for texts rooted in cultural or philosophical domains.This project differs by fusing translation with meaning-based retrieval. By using Sentence Transformers for semantic understanding and FAISS for high-speed vector comparison, the tool adds a new dimension to multilingual NLP. It also uniquely features a domain-curated Indic corpus and optional BLEU evaluation — features not commonly found together in mainstream systems.

|  |  |  |
| --- | --- | --- |
| **Tool** | **Features** | **Limitations** |
| **Google Translate** | Language detection, high-quality translation, many language pairs | No semantic search; no BLEU; no download; closed-source |
| **DeepL** | High-quality translations (esp. Euro-centric), language detection | No BLEU; no domain context; limited language range (no Sanskrit, etc.) |
| **Facebook NLLB Demo** | NLLB-based translation for 200+ languages | No semantic context; no BLEU; no report export |
| **OpenNMT / Fairseq Pipelines** | Research-grade translation pipelines | Requires heavy setup; no GUI; not user-friendly |
| **Bing Translator / Amazon Translate** | Commercial APIs for translation | No BLEU, similarity search, or cultural context |
| **Sanskrit Heritage Reader / SANSK-NLP** | Sanskrit tokenization and translation | Focused only on Sanskrit; limited generalization; no multilingual support |

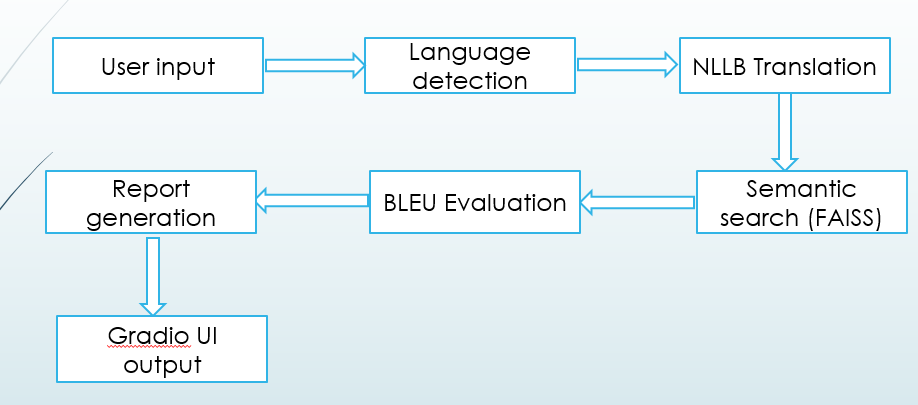
# 

# System Design

The system operates as a pipeline. First, the input text is analyzed by a language detection model based on XLM-RoBERTa. The detected language code is mapped to a compatible token for the NLLB-200 translation model, which produces the translated sentence in the target language.

Next, the translated sentence is embedded into a vector using the SentenceTransformer model. This vector is compared against a FAISS index containing pre-computed embeddings from a curated corpus. The top 3 semantically closest matches are retrieved.

If a human translation is provided, the system computes a BLEU score using the SacreBLEU library, offering a measure of translation accuracy. All outputs are plotted (semantic match scores) and compiled into a downloadable report.



The entire process is packaged in a Gradio-based interface for accessibility.

# Implementation Details

The project is implemented in Python using several core libraries:

* Hugging Face Transformers for language detection and NLLB translation
* Sentence Transformers for vector embeddings
* FAISS for similarity search
* SacreBLEU for BLEU score computation
* Matplotlib for plotting
* Gradio for the web interface

The architecture involves loading models at runtime, preprocessing inputs (tokenization, truncation), computing vector representations, and storing corpus vectors in memory. The static corpus is encoded once at initialization. Results are written to a temporary text file which can be downloaded by the user.

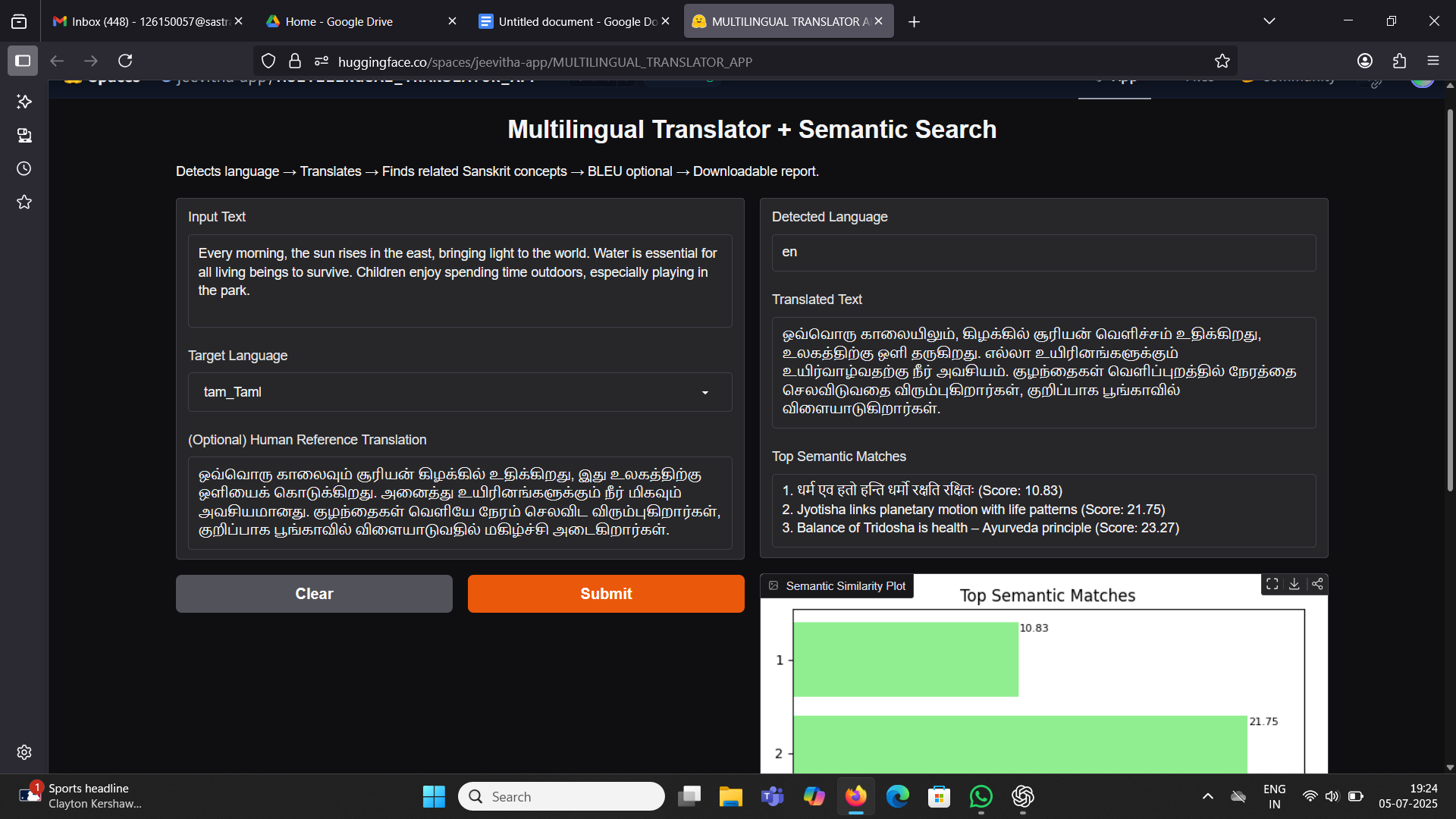
# 

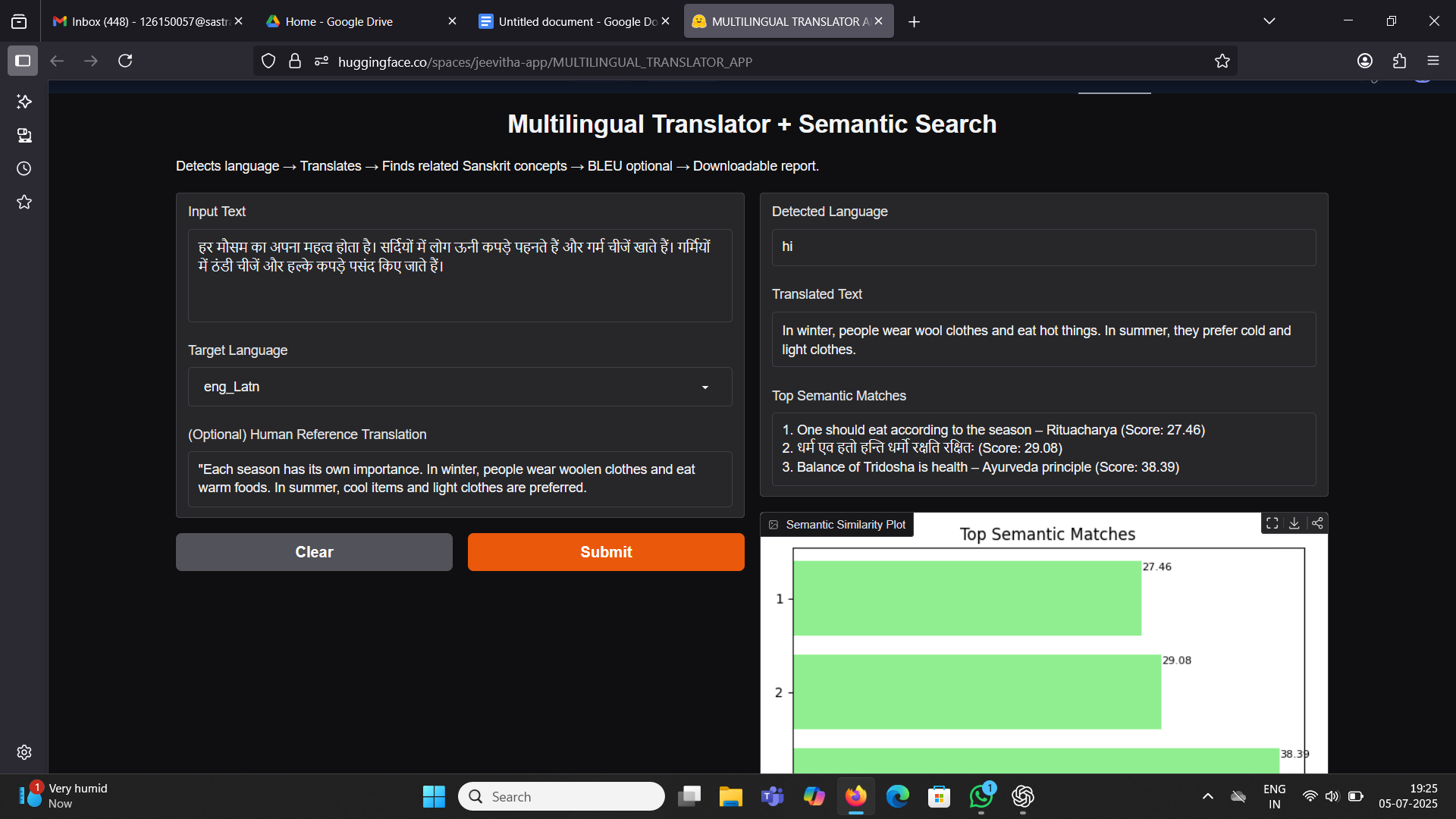
# Results and Analysis

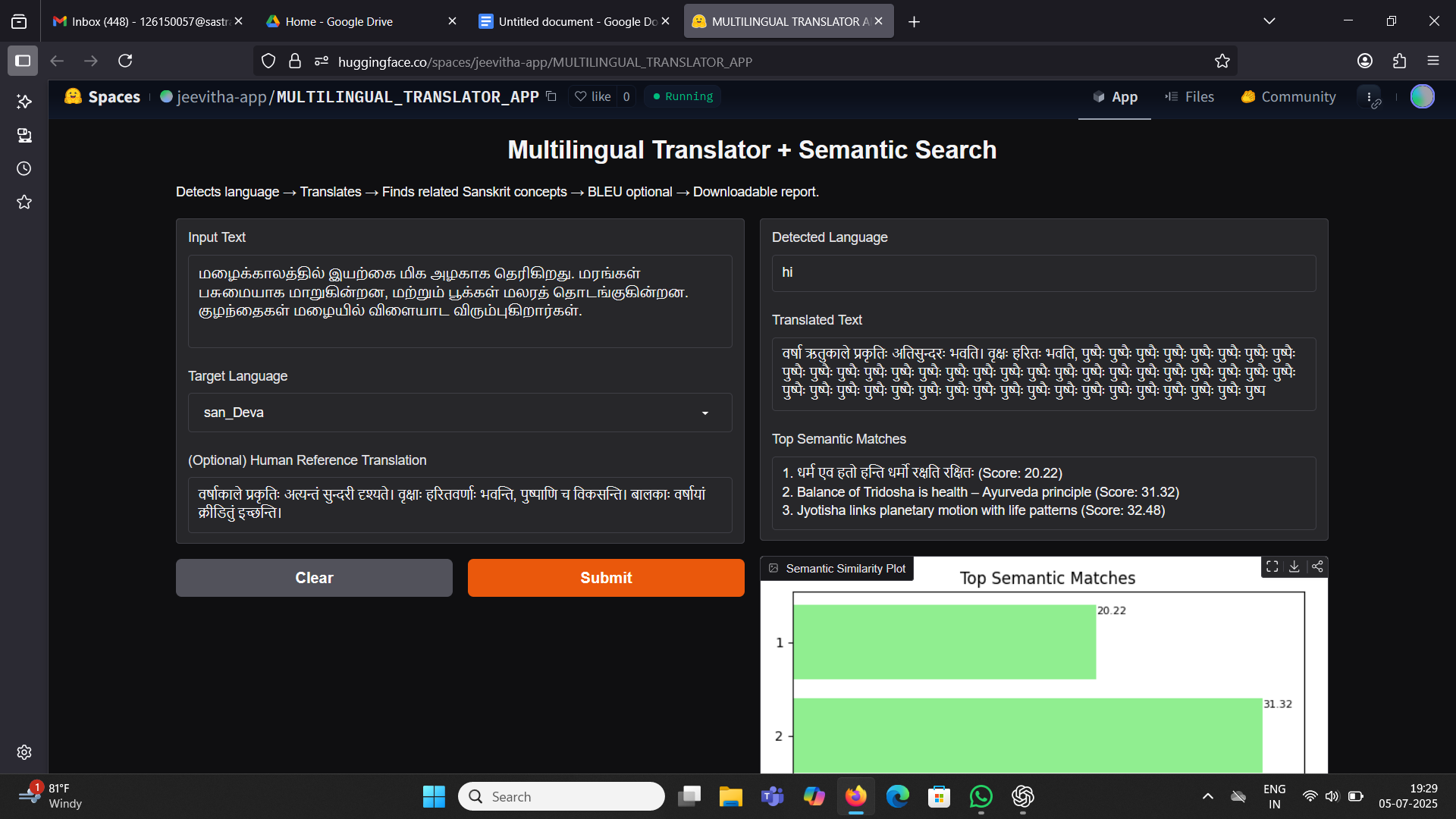
The system accurately detects languages including Hindi, Sanskrit, Tamil, Chinese, and Arabic. Translations into English or other target languages are fluent and meaningful, especially when translated from supported low-resource languages.

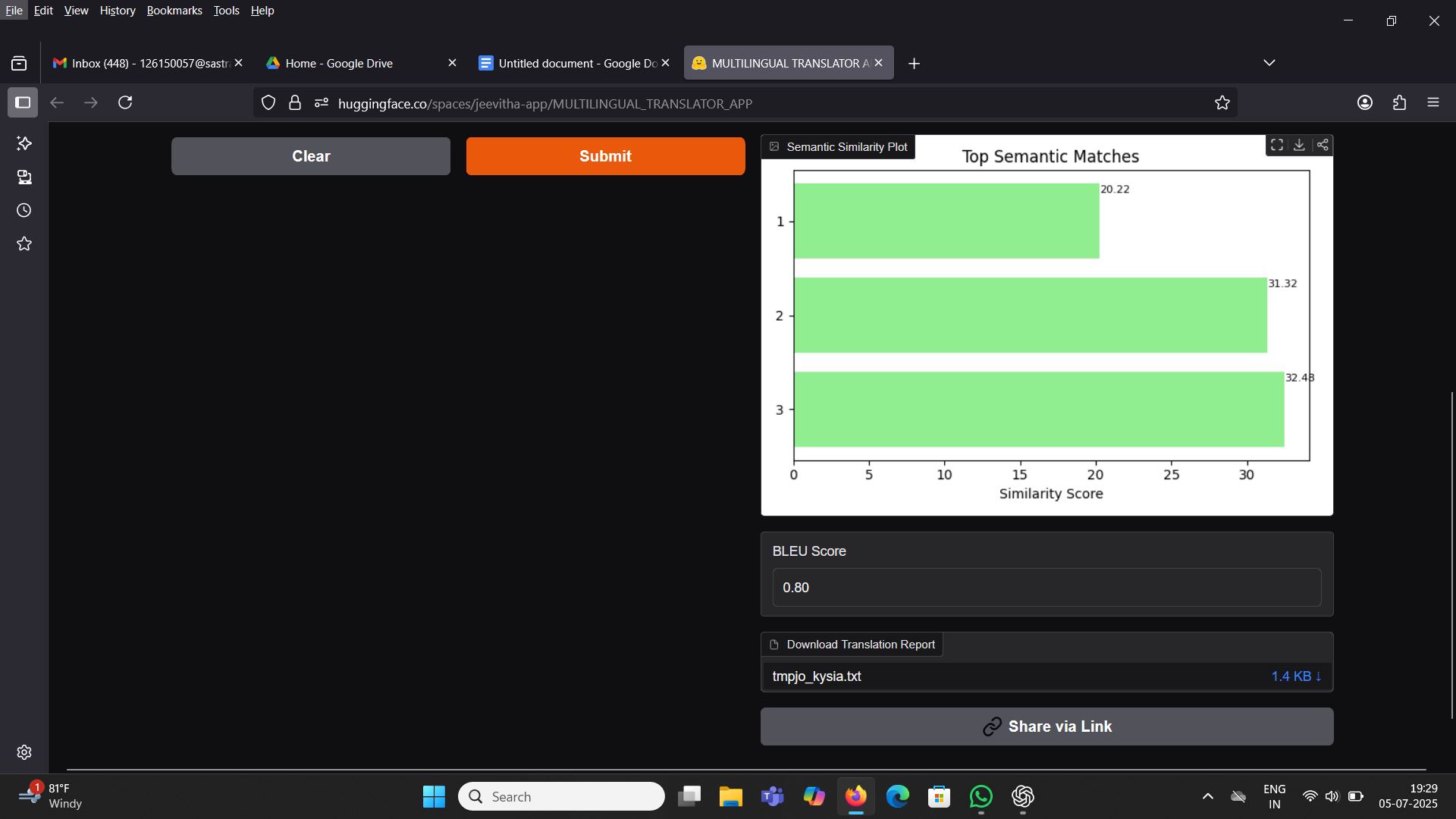
Semantic search consistently retrieves related Indic or philosophical concepts. For example, the translated text "The balance of body energies leads to health" returns corpus matches related to Ayurveda and Tridosha.

The BLEU score provides quantitative insight when comparing to human references. Semantic similarity scores are plotted and display meaningful variation based on context.









# Comparison with Existing Tools

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Google Translate** | **DeepL** | **This Project** |
| Language Detection | Yes | No | Yes |
| Low-resource Language Support | Moderate | Low | High (via NLLB-200) |
| Semantic Search | No | No | Yes (via FAISS) |
| BLEU Score Evaluation | No | No | Yes (Optional) |
| Cultural Corpus Matching | No | No | Yes (Indic corpus) |
| Downloadable Reports | No | No | Yes |

This project stands out by not just translating, but connecting meaning with culturally aligned knowledge — an improvement over syntax-only systems.

# 

# Key Innovations and Distinctive Features

While many tools offer language detection or translation in isolation, this project stands out by integrating multilingual NLP with cultural intelligence and semantic understanding - all in a single, cohesive interface.

Unlike commercial systems such as Google Translate or DeepL, which prioritize speed and general-purpose translation, this tool is crafted with a deeper interpretive goal. It not only translates - it contextualizes. By connecting translated output to a curated corpus of Sanskrit-based philosophical and scientific ideas, the system helps users discover how modern language maps to ancient wisdom.

Furthermore, this tool goes beyond surface-level translation. It embeds both input and output into a semantic vector space, enabling users to explore meaningful connections between languages and concepts. This feature is not available in any commercial or open-source translation platform today.Another rare inclusion is the optional BLEU evaluation. Users can compare the machine translation against human reference translations, providing a transparent measure of quality -an essential feature for researchers, educators, and linguists.

Lastly, the application is designed for **clarity and usability**. It offers:

* An intuitive web interface (Gradio)
* Clear visual plots for similarity scores
* A downloadable report for documentation or review

Together, these make it more than just a translator — it's a cultural and semantic interpreter, uniquely suited for interdisciplinary applications in AI, linguistics, Sanskrit studies, education, and cross-cultural communication.

# Conclusion

This project successfully integrates multilingual translation with semantic search, enriching the translated output with relevant cultural context. It solves a key problem in NLP: bridging language with meaning, especially in culturally rooted content. The BLEU score and downloadable reports add value for academic and research-oriented users.

The modular structure and Gradio interface make the tool extensible and easy to use. Despite limitations like static corpus size and reliance on pre-trained models, the system demonstrates practical value in combining translation with understanding.

# Future Work

* Dynamic corpus upload support
* Dynamic input support system
* Combined interface (file + text input)
* User-adjustable number of semantic results (top\_k)
* Improved error messaging and edge-case handling
* Hindi/Sanskrit-specific quality filters
* REST API backend for programmatic access

# Reference

[Multilingual-translator app](https://huggingface.co/spaces/DheivaCodes/Multilingual-translator)