2024 AI Challenge – BasicLingua AI Powered NLP

Team Name: Zevar

Product

Problem Statement

The amount of text data in the world is growing rapidly, with an estimated 40 zettabytes expected by 2024 [1]. This rapid growth creates challenges for natural language processing (NLP) models, which can struggle with complex language patterns, leading to error rates between 10% and 30% [2]. NLP libraries that offer solutions are either limited in their ability to solve the required problem or require a great deal of human intervention to handle text data. To advance NLP, future work should focus on managing this data explosion, improving accuracy, reducing manual effort, and closing the language gap to make NLP more inclusive.

Our Solution

Our BasicLingua NLP library, built for Python programming, tackles the challenges of accuracy and effectiveness in NLP. It uses generative AI (Large Language Models) to bridge the language gap and enables the developers to solve NLP tasks through natural language prompts. Since it's based on generative AI, BasicLingua supports a wide range of NLP tasks for with improved accuracy and results. This reduces the need for extensive human involvement.

Target Community and Reach

BasicLingua is designed for anyone who works with text data, including:

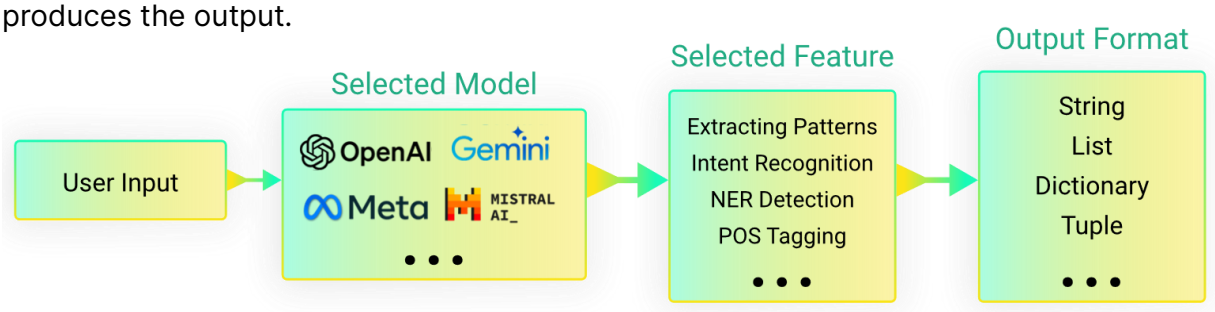
- Software Developers: Integrate advanced NLP features into applications with ease.
- Data Scientists: Analyze and extract insights from large text datasets efficiently.
- Researchers: Explore new possibilities in NLP research with a powerful toolkit.
- Educators: Enhance teaching and learning experiences with innovative text analysis tools.

The library is user-friendly design and AI Powered documentation make it accessible to users with varying levels of technical expertise.

Engineering

High-level Design

BasicLingua library is easily accessible through `pip install`, like any other Python library. It takes user input and, based on the selected model and requested feature, it produces the output.



BasicLingua provides two types of documentation to boost productivity. There's manual documentation you can access through any code editor while coding. We also have an AI-powered guide that customizes its features to match your task, saving you time and effort.

Role of Generative AI

More than 90% of our library relies on generative AI. We've used the most prominent Large Language Models as the backend for our library to process user data. Prompt engineering techniques play a significant role in making this system more cost-efficient by ensuring it outputs only the most relevant information, with a limited number of words, to minimize costs.

Evaluation

Our library was tested on diverse datasets, including clinical notes from MIMIC-IV, Wikipedia articles, and AI-generated text. This range allowed us to assess features like entity extraction and intent detection across different contexts.

Task Name	Evaluation Metric	OpenAI (GPT-3.5)	Gemini (Gemini-1.0)	AnyScale (Llama-3-70b)
Information Extraction	Average F1 Score	0.85	0.64	0.77
Analysis	Average Accuracy	0.90	0.74	0.81
Summarization	ROUGE Score	0.78	0.68	0.72
Coreference & Disambiguation	Average F1 Score	0.89	0.73	0.84
Processing	Average Accuracy	0.88	0.73	0.82

Going Forward

With its success, BasicLingua will continue to evolve by:

- Expanding LLM Support: Integrate new and emerging LLMs to offer even greater flexibility and choice.
- Enhancing Functionalities: Develop new features and refine existing ones based on user feedback and evolving NLP needs.
- Building a community: Collaboration and knowledge sharing among users through forums, tutorials, and other initiatives.
- Exploring Multimodal Capabilities: Integrate text and vision capabilities to unlock new applications in areas like image analysis and document understanding.

References:

1. IDC. (2022). Global Datasphere. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=IDC_P38353

2. Li, N., Noh, S., & Liu, Y. (2021). CSMD: Context-Sensitive Multi-Task Detection of Sarcasm or Deception in Social Media Text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 2789–2802). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X20312926