Natural Language Processing Project Proposal – Group 2

Project Members: Sairam Venkatachalam | Devarsh Sheth | Disha Kacha | Smit Pancholi

Project Topic:

For our project, we aim to develop a chatbot for interacting with historical leaders from different countries. This project focuses on creating an interactive tool that allows users to ask questions and compare responses from leaders across different eras, and explore perspectives in multiple languages. By leveraging NLP techniques for sentiment analysis, machine translation, and dialogue generation, this chatbot aims to make political history more engaging and accessible.

Data source:

We will utilize publicly available APIs to obtain U.S. presidential speeches and access translated archives for speeches by German and Russian leaders. Additionally, we have requested data from relevant archives to include speeches by Indian leaders, ensuring a comprehensive dataset.

Our team will gather, standardize, and categorize the speech data to enable easy access and comparison. German and Russian speeches will be translated into English, supporting a unified analysis across all sources.

To address translation nuances and potential data gaps, especially for non-English speeches, we will fine-tune our models and, when necessary, supplement missing information with reliable secondary sources.

NLP Frameworks and Packages:

We plan to use the following methods in the NLP domain:

- **Data Preprocessing and cleaning:** We will clean and structure multilingual speech data for consistency. This would include the use of the NLTK and spacy packages for cleaning data
- **Machine Translation:** German and Russian speeches will be translated to English with a fine-tuned model to retain context.
- **Text Generation:** A model will be trained to simulate responses in each leader's unique language style. This will be done using a combination of a next token prediction model, and if necessary, a Reinforcement Learning with Human Feedback (RLHF) step in order to enable the chatbot to answer questions
- **Sentiment Analysis:** Sentiment and topic modeling will identify key themes and emotions in speeches. This could be used to analyze differences in opinions between eras and countries
- **Retrieval Augmented Generation:** In order to generate better responses, we might also need to implement an information retrieval RAG system, to get the relevant context from the speeches

As detailed above, our project will cover the NLP tasks of **Text-Generation**, **Machine Translation** and **Sentiment Analysis**

Performance Evaluation:

The evaluation phase will involve testing translation accuracy, conversational flow, and sentiment extraction to measure the chatbot's effectiveness in delivering relevant, contextually accurate responses. We expect the chatbot to perform well in these areas, enhancing user engagement with historical insights and ideologies.

For the RAG process, we could also use TRULens to evaluate the groundedness and context relevance

Project Schedule:

- Week 1: Data preprocessing and Cleaning
- Week 2-3: Designing and training the Chatbot with the prepared data
- Week 4: Model Evaluation, fine-tuning and RAG implementation
- Week 5: Creating interactive application and project finalization

This is the rough schedule we have in mind.