**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. By leveraging NLP techniques for sentiment analysis, summarization, and named entity recognition, this chatbot aims to make political history more engaging and accessible.

**Data source:**

We will utilize publicly available APIs and archives to gather a dataset consisting of speeches by U.S. and Russian presidents. The U.S. presidential speeches will be collected from official government websites and public archives, while the Russian dataset will be obtained from reliable online repositories, already available in English. Our team will standardize and categorize the speech data to ensure consistency and enable seamless analysis and comparison across both sources.

**NLP Frameworks and Packages:**

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

- **Data Preprocessing and Cleaning**: Speech data will be cleaned and structured for consistency using NLTK and spaCy.
- **Text Generation**: A fine-tuned model will simulate responses in the unique rhetorical style of each leader. This will include generating text based on next-token prediction, with potential integration of summarization and named entity recognition (NER) to enhance the relevance and informativeness of responses.
- **Sentiment Analysis**: Sentiment analysis will identify emotions and key themes in speeches, enabling insights into differences in tone and focus across eras and countries.
- **Retrieval-Augmented Generation (RAG):** To improve response quality, we will implement a RAG system using ChromaDB for efficient data retrieval and Sentence Transformers for embedding generation. This will ensure that responses are grounded in the most relevant speech excerpts.

As detailed above, our project will cover the NLP tasks of **Text-Generation**, **Summarization, Named Entity Recognition** and **Sentiment Analysis.**

**Performance Evaluation:**

The evaluation phase will focus on a human-in-the-loop approach to iteratively refine the chatbot's performance. By incorporating human judgment and feedback into the process, we can assess and improve the chatbot's ability to deliver relevant, contextually accurate, and user-friendly responses. This approach allows for a more practical and qualitative evaluation, emphasizing conversational flow, tone, and relevance over traditional metrics.

The human-in-the-loop method is essential for fine-tuning the chatbot to align with user expectations and address any shortcomings dynamically. It ensures continuous improvements, making the chatbot more effective in engaging users with historical insights and ideologies while maintaining clarity and precision in its responses.

**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.