Group 9

Speech Impact Analyzer

1. Choice of Dataset

For this project, we plan to put together a mixed dataset of political speeches and tweets from public figures so we can compare how rhetorical choices change between longer, prepared addresses and short, online messages

The dataset will include:
- Speeches: official addresses by politicians or world leaders, with metadata such as speaker, date, country, and topic.
- Tweets: posts from the same or similar figures with engagement metrics (likes, retweets, replies).

Our goal is to study how logos (facts and reasoning), pathos (emotion), and ethos (credibility) show up in each format and whether those patterns relate to audience reaction.

Engagement levels will be labeled as High, Medium or Low, based on relative interaction counts. For speeches, engagement proxies such as view counts or like ratios from official channels will be used when available. The dataset will be limited to English-language content, cleaned for consistency, and balanced across speakers and time periods.

2. Methodology

a.

The text will be cleaned and processed using basic NLP techniques. We will remove URLs, emojis, and special characters, convert text to lowercase, tokenize words, and remove stopwords while retaining punctuation that conveys meaning (!, ?).

We will extract rhetorical and linguistic features associated with logos, pathos, and ethos:
- Logos: numbers, logical connectors (“because”, “therefore”), and factual or analytical language.
- Pathos: emotional vocabulary, exclamation marks, rhetorical questions, and pronouns (“we”, “you”).
- Ethos: authority markers (“as a leader”, “experts”, “trust”) and confident modal verbs (“will”, “shall”).

Each text will receive sentiment and subjectivity scores, and feature frequencies will be normalized to produce scores between 0 and 1 for logos, pathos, and ethos.

b.

We plan to train a supervised classification model that predicts audience engagement levels (High**,** Medium, Low) based on the rhetorical and linguistic features we extract from the texts. In simple terms, the model will learn which types of expressions, tones, or rhetorical choices are most often linked to higher engagement. To make sure the evaluation is fair, we'll split the data into training and testing sets and make sure that speeches or tweets from the same speaker don't appear in both.

We'll start with straightforward and interpretable algorithms, such as logistic regression or decision trees, because they're easier to explain and require less tuning. Once we get a baseline, we'll see if adding more complex models actually improves the results. The model will output two main elements: the predicted engagement level with its confidence score, and the relative balance of logos, pathos, and ethos, so we can interpret which rhetorical dimensions seem to influence engagement the most.

c.

We will evaluate the model's performance using standard metrics such as accuracy, precision, recall, and F1-score to measure how well it predicts each engagement category. A confusion matrix will help us see where the model performs well and where it struggles. We will also compare our results to a simple baseline that always predicts the majority class, to confirm that our model provides a real improvement over naive guessing.

3. Application

The goal of the project is to develop a simple web application that analyzes and compares rhetorical styles in political speeches and tweets.

User Input: Users can paste a tweet or a short excerpt from a speech. The system will process the text, extract rhetorical features, and estimate how similar language typically affects engagement.

Model Output: The app will display a rhetorical profile (logos–pathos–ethos scores), a predicted engagement level (High, Medium, Low), and a short interpretation such as: "This text uses strong emotional language and moderate logical structure, similar messages tend to generate high engagement."

Purpose: This application bridges AI and classical rhetoric, showing how persuasion adapts to both traditional and digital communication. It provides an educational tool to visualize the art of eloquence through measurable data.