Group 9
Speech Impact Analyzer

1. Problem statement

The Speech Impact Analyzer project investigates how rhetorical choices, (specifically logos (reasoning), pathos (emotion), and ethos (credibility)), influence audience engagement across political speeches and tweets. Our goal is to compare how persuasion changes between formal addresses and short online messages, using engagement metrics such as views, likes, and retweets as proxies for audience reaction.

2. Data Preprocessing

For this project, we worked with two Kaggle datasets: *"EMPOLITICON – NLP and ML Based Approach for Context and Emotion Classification of Political Speeches"* and the *"Political Tweets Dataset."* The first contains approximately 7,000 political speech transcripts with metadata such as speaker, country, and emotion labels, while the second includes around 12,000 tweets from verified politicians along with engagement metrics like likes, retweets, and replies. We initially planned to build our own dataset, but we chose these Kaggle sources for their reliability, structured format, and availability of relevant features, which allowed us to focus on analysis rather than manual data collection. Both datasets were filtered to include only English content and cleaned to remove duplicates, missing entries, and irrelevant characters such as URLs, emojis, and mentions. The text was converted to lowercase, tokenized, and lemmatized, while stopwords were removed except for rhetorically significant ones like "we" or "you." Engagement levels were categorized as *High*, *Medium*, or *Low* based on quantile thresholds of likes, retweets, or view counts, ensuring a balanced distribution across classes. We retained punctuation marks like exclamation points and question marks to preserve emotional and rhetorical cues. From each text, we extracted linguistic and rhetorical indicators for *logos* (reasoning and factual language), *pathos*(emotional and affective tone), and *ethos* (credibility and authority markers), normalizing these scores between 0 and 1 for comparability. Finally, the idea is ultimately to have both datasets merged into a unified structure containing the fields: text, speaker, source type (speech or tweet), logos score, pathos score, ethos score, and engagement level. The final preprocessed dataset would include approximately 19,000 samples and forms the foundation for training and evaluating our engagement prediction model.

3. Machine learning model

We want to develop a baseline classification model to predict audience engagement levels—Low, Medium, or High, based on rhetorical and linguistic features. The model would be implemented in Python using the scikit-learn library, with additional tools such as pandas, matplotlib, and NumPy for data manipulation and visualization. Our pipeline combines TF–IDF text representations with handcrafted rhetorical indicators (logos, pathos, ethos) and passes them into a multinomial Logistic Regression classifier. This design was chosen for its interpretability, computational efficiency, and strong baseline performance on textual data.

We selected L2 regularization to prevent overfitting and optimized the model using the SAGA solver, which handles sparse TF-IDF matrices efficiently. The dataset was split into 80% training and 20% testing, stratified across engagement levels to ensure balance. Hyperparameters such as the regularization strength (C) and maximum number of features for the TF-IDF vectorizer were tuned through cross-validation.

To validate our approach, we employed stratified 5-fold cross-validation on the training data, measuring accuracy, precision, recall, and F1-score for each engagement class. We planned to visualize performance using a confusion matrix, learning curves, and feature importance plots (highlighting influential words and rhetorical features). During implementation, we will probably encounter challenges related to incomplete engagement data in the tweet dataset, which temporarily prevented multi-class label generation. This issue will be addressed by integrating a complete version of the dataset or by recalibrating engagement thresholds per source type. Once resolved, we will finalize training and save the model weights, TF-IDF vectorizer, and scaler for deployment in the Speech Impact Analyzer system.

4. Preliminary results

In early runs of the model on partial data, we hope that the data will show promising differentiation between engagement levels, particularly for speech transcripts where emotional tone (pathos) and credibility cues (ethos) strongly align with high engagement. The model achieved balanced accuracy across classes, though tweets displayed more variance due to shorter text length and missing engagement metrics. We plan to refine these results with additional data and visual analysis of rhetorical weightings.

5. Next steps

Moving forward, we will (1) complete data labeling for tweets to enable consistent multi-class training, (2) fine-tune hyperparameters using grid search, and (3) explore more sophisticated architectures such as BERT or XGBoost for deeper contextual understanding. These models will capture subtler rhetorical nuances while maintaining the interpretability of our rhetorical scores. Ultimately, this iterative approach will improve predictive accuracy and strengthen the model's ability to reveal how logos, pathos, and ethos influence audience engagement across political communication formats.