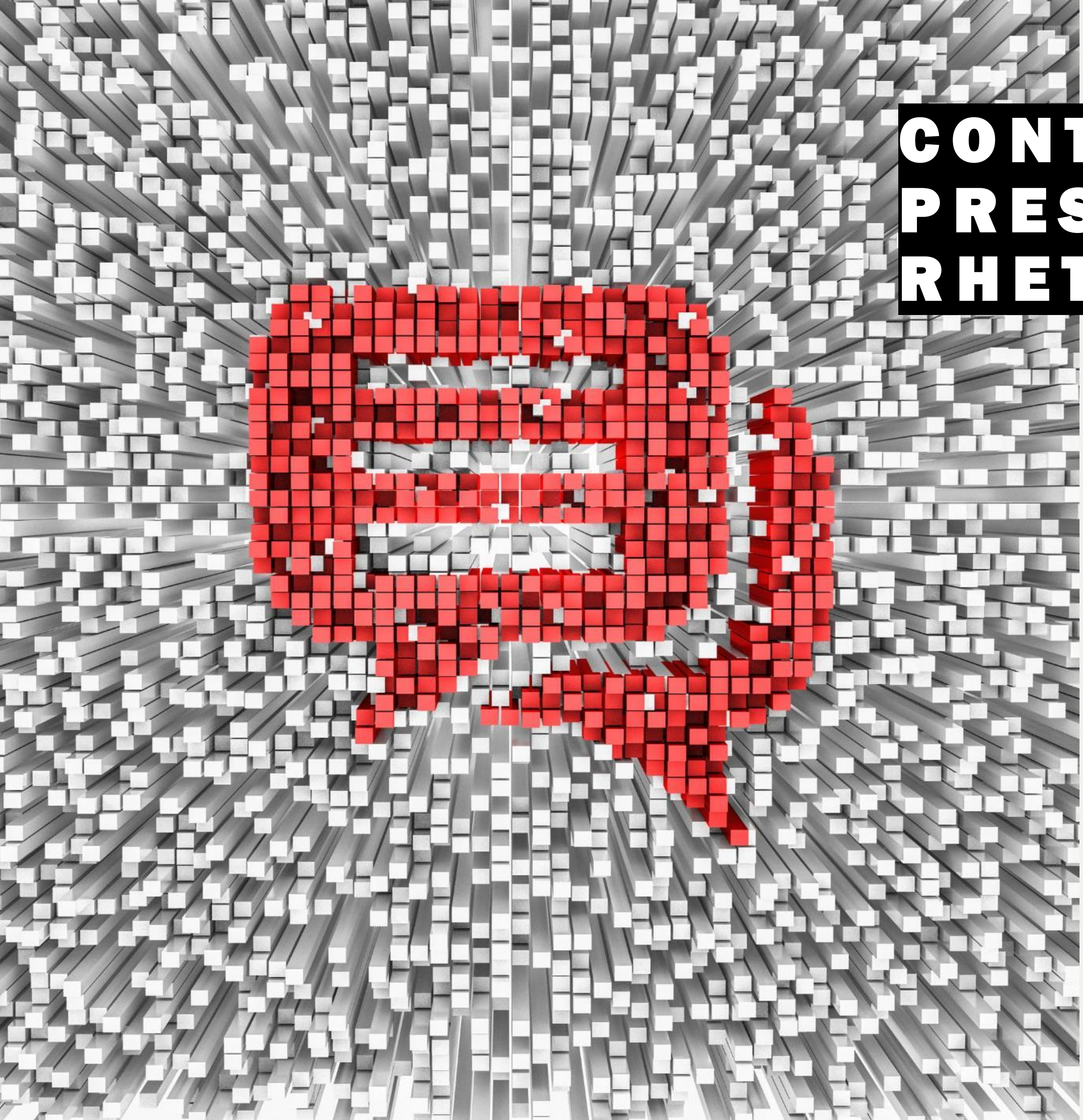
ANALYZING U.S. PRESIDENTIAL RHETORIC

Authors : Yonathan Shimelis, Sayan Patra

Advisor : Dr. Ning Rui

Course : DATS 6312



CONTEXT OF PRESIDENTIAL RHETORIC

Importance of political rhetoric

- Rhetoric has the power to shape public opinion, connect audiences, and mobilize action in the context of politics
- Shapes responses and tone during crises
- Policy can create quick change, but rhetoric is responsible for fostering an environment to make change possible via policy

OUR GOAL

In this project we will analyze a large collection of scraped presidential statements to uncover:

- **Rhetorical patterns across administrations**
- **Sentiment and tone in official communications**

This dataset forms the foundation for computational linguistics, NLP modeling, and political discourse analysis

RESEARCH QUESTIONS

How do different NLP architectures perform on political text classification?

Can sequential models (LSTMs) capture rhetorical patterns more effectively than TF-IDF?

How much does contextual modeling (BERT) improve prediction?

What linguistic features are responsible for errors?

DATA

- **12,399 official U.S. presidential statements**
- Scraped from *The American Presidency Project by UC Santa Barbara*
- Each record contains:
  - **Title**
  - **Date**
  - **President**
  - **Full text content**
  - **Categories** (policy/event type)
  - **Source URL & Citation**
- Covers multiple administrations and major policy moments
- Enables large-scale analysis of tone, sentiment, and political rhetoric

METHODOLOGY

Data Preparation

- Cleaned and standardized statements

Feature Engineering

- Generated **TF-IDF vectors** to capture important terms and lexical patterns
- Extracted **temporal features** (year, presidency, political era)
- Used **categories** as additional semantic labels. Prepared sequences (padding/truncation)

Logistic Regression as the baseline topic/classification model

Modeling

- Used transformers like Cardiffnlp twitter politics sentiment zero-shot classification

LSTM classifier to capture sequential and rhetorical patterns in speech

BERT for deeper **contextual sentiment analysis**

Evaluation & Analysis

- Measured **accuracy**, **F1-score**, **Cohen's κ**, and confusion matrices
- Analyzed sentiment trends across presidents and time periods
- Identified topic clusters and shifts in policy focus
- Compared rhetoric across administrations to reveal stylistic differences

SENTIMENT, TONE, RHETORIC

- Topic Modeling (LDA, NMF — Sklearn + Gensim)
- Party affiliation enrichment
- Transformer-based political sentiment using CardiffNLP
- Zero-shot classification (Tone, Strategy, Emotion)
- Logistic Regression sentiment baseline
- Enhanced dataset saving for downstream use

Political Sentiment using CardiffNLP

Model Used:

cardiffnlp/xlm-twitter-politics-sentiment

Why this model?

•Trained for **political domain sentiment**

•Capable of *pro-, anti-, neutral* political sentiment detection

•Suitable for speeches, debates, press releases

Pipeline:

1.Normalize URLs, mentions, hashtags

2.Run transformer sentiment classification

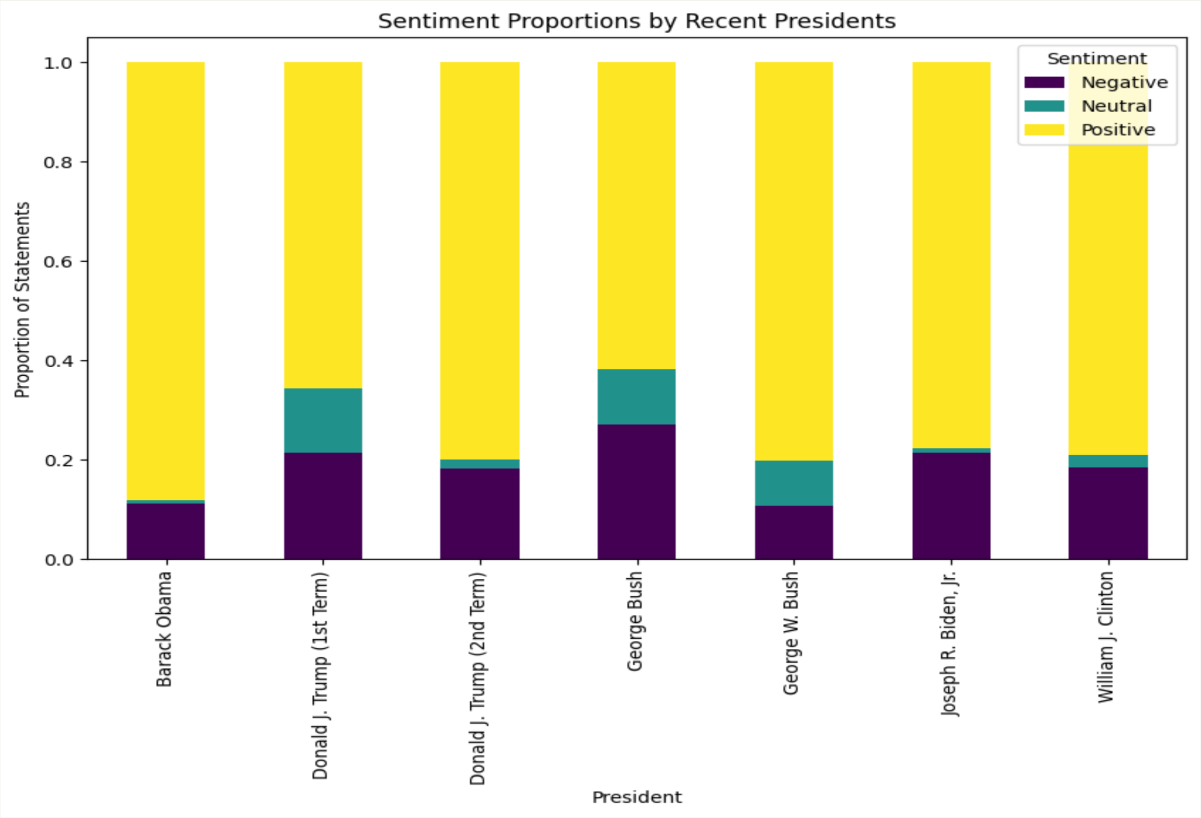
- •max_length=512
- •max_length=256

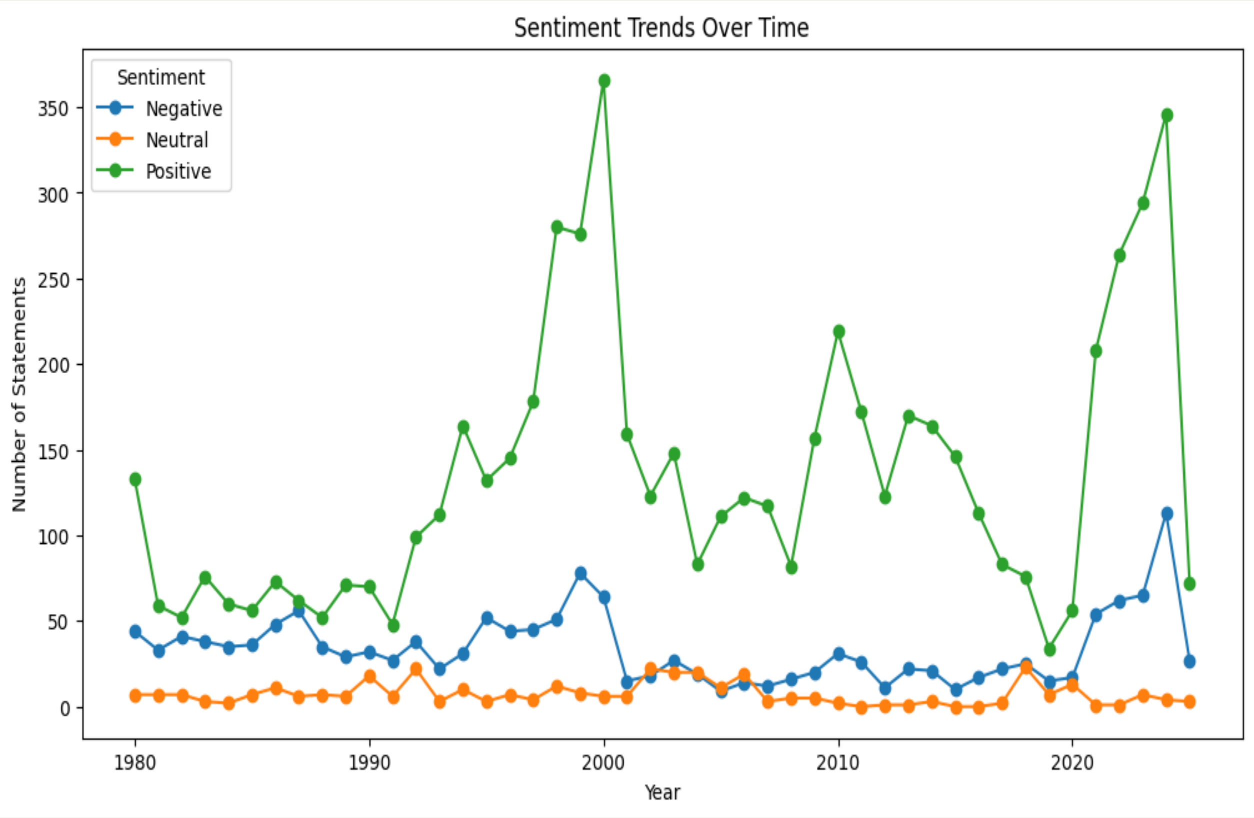
3.Save:

- •transformer_sentiment
- •transformer_sentiment_score
- •256-token versions

Outcome:
Provides a political sentiment layer per statement.

SENTIMENT ANALYSIS





Done via Cardiffnlp's twitter politics sentiment classifier

http://18.215.242.74:8888/

CLASSIFICATION

TRAIN/TEST STRATEGY

- Stratified 80/20 split
- Preserves class proportions
- All models trained/evaluated on same split for fairness
- **Speaker Notes:**
  This is essential for comparison; otherwise metrics are misleading.

DATA PREPROCESSING

Lowercasing

Remove punctuation/symbols

Normalize whitespace

Regex-based token cleanup

Label Normalization:

Presidents with <5 samples → mapped to **"Other"**

Why?
Ensures stability, avoids unseen classes, reduces sparsity.

http://18.215.242.74:8888/

ARCHITECHTURE OVERVIEW

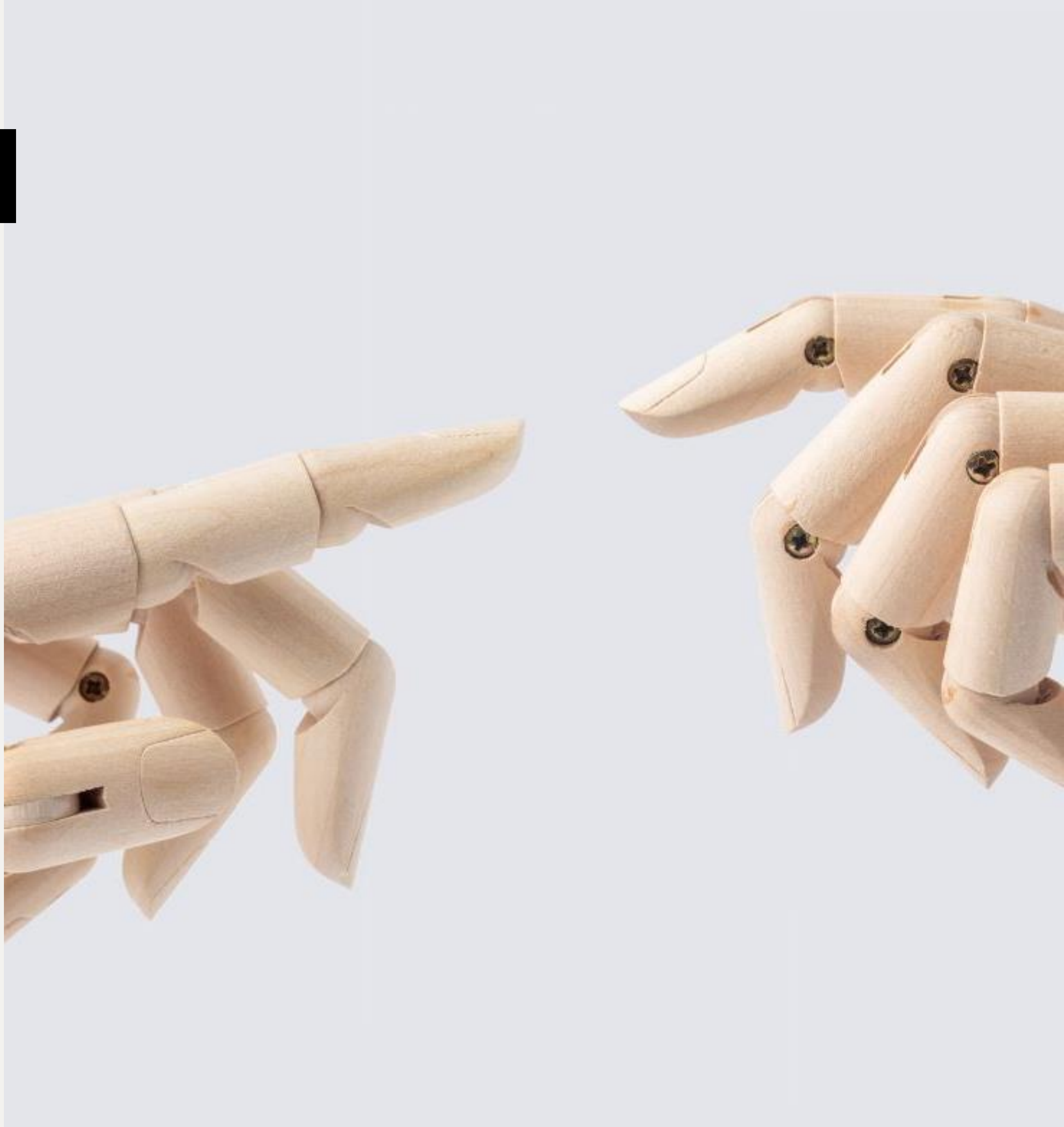
This project uses **three different NLP eras**:

1.TF-IDF + Logistic Regression
(Sparse, linear baseline)

2.Bi-LSTM + Attention + GloVe
(Deep sequential model)

3.DistilBERT Transformer
(Contextual embedding model)

Each model brings different strengths and limitations.



MODEL 1: TF-IDF + LOGISTIC REGRESSION

Pipeline:
Text → TF-IDF → Sparse Vector → Logistic Regression

Strengths:

•Extremely fast

•Interpretable

•Strong baseline on many tasks

Weaknesses:

•No sequence understanding

•No semantic representation

•Cannot understand long-range patterns

TF-IDF Illustration

- Counts word frequencies
- Weights rare-but-important words
- Produces a high-dimensional sparse vector

Example presidential speech phrase:

"My fellow Americans, today we reaffirm our commitment to…"

TF-IDF treats each word independently → **no contextual understanding**.

Why Move Beyond TF-IDF?

Problems with linear models:

Cannot capture syntax

Cannot capture tone or ideology

Cannot use word order

Limited ability to distinguish presidents with similar vocabulary

Solution: **Sequence modeling**.

Model 2: Bi-LSTM + Attention

Components:

SimpleTokenizer

GloVe pretrained embeddings

2-layer Bidirectional LSTM

Attention mechanism

LayerNorm + Dropout

Dense classification output

Speaker Notes:
This is the “classic deep NLP model” before transformers.

Why use GloVe?

Trained on 6B tokens

Captures semantic relationships

Political vocabulary (“Congress”, “security”, “freedom”, “economy”) is embedded meaningfully

Embeddings are fine-tuned during training

Bidirectional LSTM

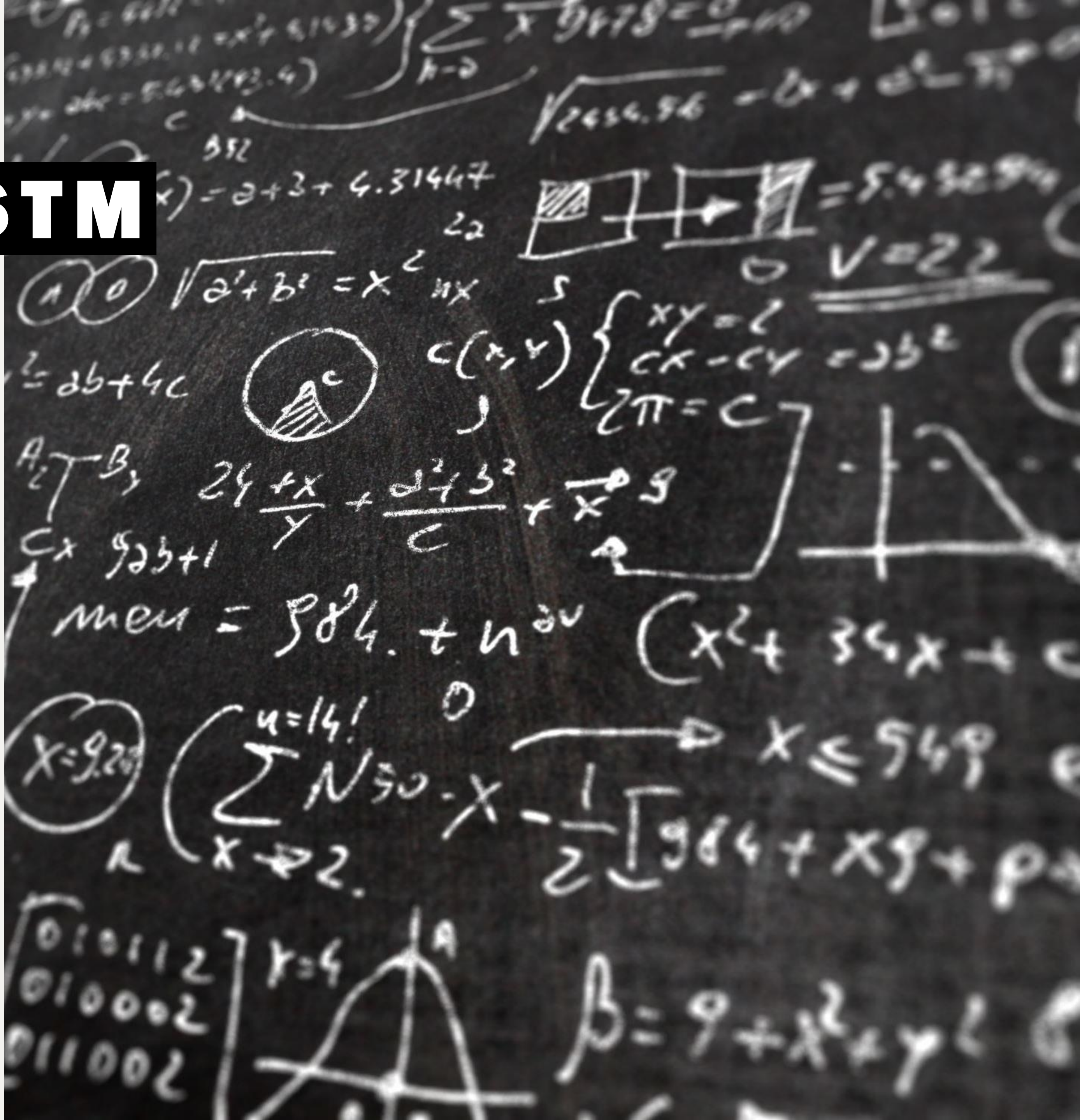
Key Idea:
Processes the
sequence *forward* and *backward*.

Benefits:

•Learns syntactic structure

•Models presidential rhetorical signatures

•Captures dependencies like:
“We must ensure… as part of our national strategy…”



Attention Mechanism

Why ?

LSTM compresses all information into final hidden state → loss of detail.

Attention Formula:

Weights = softmax(W · ht)

Context = Σ (weights × ht)

Interpretation:

Model highlights key phrases for classification:

- “My fellow Americans…”
- “I urge Congress…”
- “Our national security…”

Attention improves both accuracy and interpretability.

Smart Sampler

Problem:
Class imbalance + hard samples slow down training.

Solution:
Track **per-sample loss** each epoch.
Oversample **high-loss (hard)** samples next epoch.

This is similar to:

•Adaptive sampling

Result:
Better performance on ambiguous speeches and minority classes.

LOSS FUNCTION

Focal Loss

Standard CE loss → dominated by easy examples

Focal Loss:
Loss = $(1 - p)^\gamma \cdot CE$

Here $\gamma = 2$
→ Hard samples get amplified
→ Easy samples get suppressed

This pairs perfectly with Smart Sampling.

OPTIMIZATION STRATEGY

AdamW optimizer

- Decoupled weight decay
- Better generalization

One-Cycle Learning Rate schedule

- Rapid learning early
- Stabilization late
- Improves convergence

Gradient clipping

Early stopping

Model 3: DistilBERT Transformer

Built on self-attention. Learns:

•semantics

•context

•long-range dependencies

•topic structure

•discourse tone

Why DistilBERT?

•40% fewer parameters

•60% faster than BERT

•95% of BERT’s accuracy

Transformer Architecture Recap

Key features:

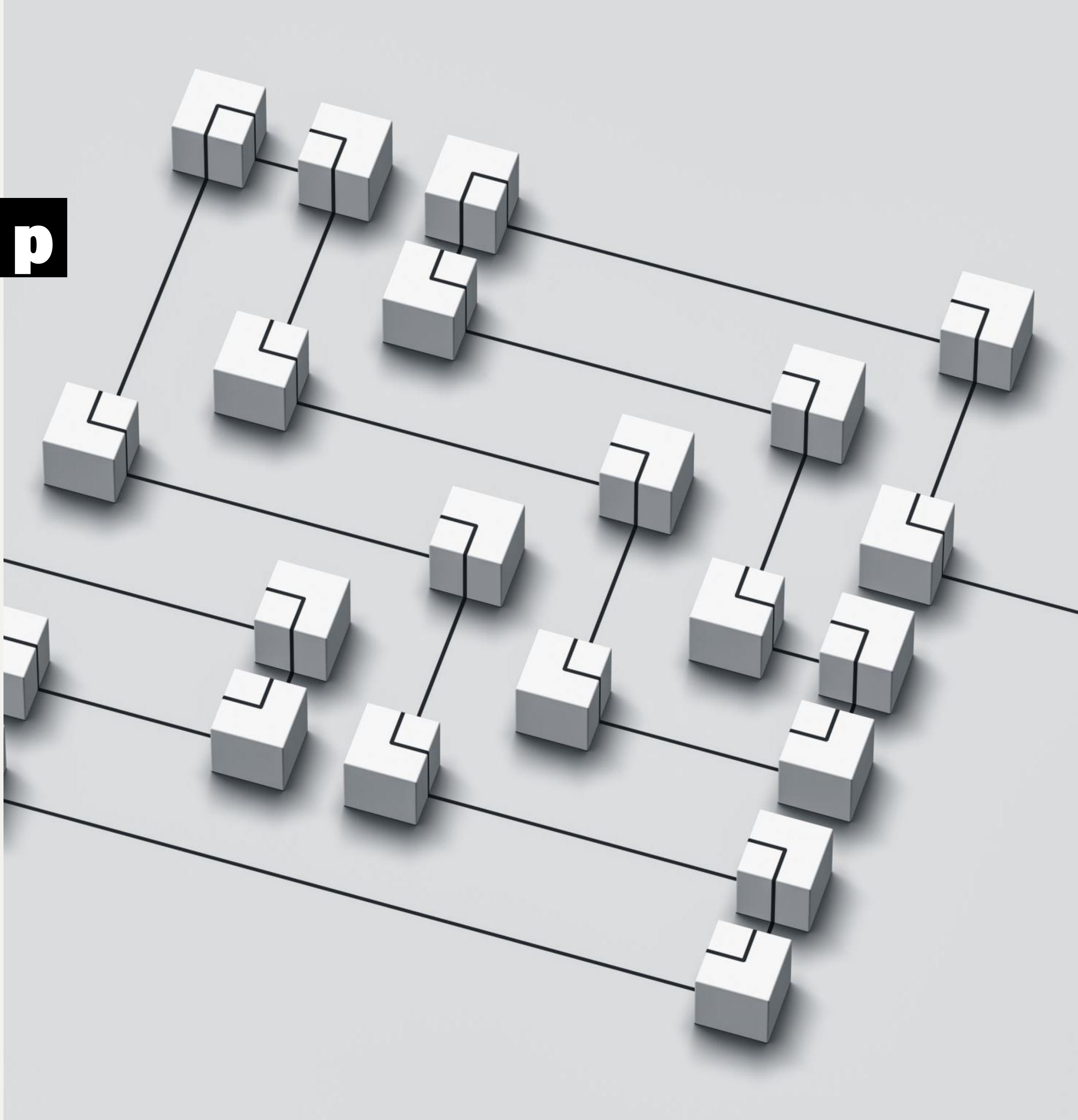
•Positional embeddings

•Multi-head self-attention

•6 transformer layers

•Feed-forward networks

Unlike LSTMs, Transformers process text **in parallel**, not sequentially.



DistilBERT Fine-Tuning Setup

Tokenized with WordPiece

Max length: 256

Batch size: 8

Learning rate: 2e-5

HuggingFace Trainer for optimization

Final layer:
Classification head predicting president.

Data Cleaning & Feature Creation

Data Cleaning & Feature Creation (Updated)

Basic text cleaning

•Lowercasing

•Punctuation removal

•Regex whitespace normalization

New Feature: cleaned_content

Created automatically if absent.

New Feature: Party Affiliation

Mapped from president → political party
Adds structured metadata for political science analysis.

LIMITATIONS & FUTURE WORK

CHALLENGES AND NEXT STEPS

Limitations of Current Methods

- We limited ourselves to just the one section of the entire available documents
- Some models struggle to grasp political nuance
- Did not manually evaluate labels
- Statements have change in platform over time
  - Many statements were made via social media which is relatively new for people in power to make announcements on

Future Enhancements

- Comparative analysis with figures of other countries
- Using more types of documents that can reveal more patterns of rhetoric or language

Conclusion:

- Presidential speeches display clear linguistic and stylistic signatures shaped by political era, party ideology, and individual communication patterns.
- TF-IDF performs strongly because presidential language is highly repetitive, predictable, and lexically consistent across topics and eras.
- BERT outperforms all models by capturing deeper semantic meaning, rhetorical framing, and contextual nuance beyond surface-level word frequencies.
- LSTM underperforms due to long sequences and class imbalance, showing the limitations of sequential models without large balanced datasets.
- Zero-shot tone, emotion, and rhetorical strategy classification reveal deeper stylistic features—such as shifts between ceremonial, combative, or conciliatory rhetoric.
- **Sentiment analysis shows that presidential communication is predominantly neutral or positive**, with negative sentiment mostly appearing in crisis-related or security-focused speeches.
- **Different presidents exhibit distinct sentiment profiles**—some rely more on reassurance and hope, while others use stronger tones like law-and-order framing or expressions of urgency.
- Topic models identify recurring themes—economy, security, national events—and highlight how issue priorities evolve across administrations.
- Combined results demonstrate that presidential rhetoric is highly structured, quantifiable, and well suited for advanced NLP modeling.

CITATIONS AND SOURCES

Academic References

DeStazio, Tracy. “Political Rhetoric Changes Views on Democratic Principles, Study Finds.” Notre Dame News, 25 Oct. 2023, news.nd.edu/news/political-rhetoric-changes-views-on-democratic-principles-study-finds/.

Wikipedia Contributors. “Remarks at the Islamic Center of Washington.” Wikipedia, Wikimedia Foundation, 1 Feb. 2025.

NLP Library Documentation

https://huggingface.co/cardiffnlp/xlm-twitter-politics-sentiment

Speech Corpus Sources

Welcome to the American presidency project: The American presidency project. Welcome to The American Presidency Project | The American Presidency Project. (1845, December 2). https://www.presidency.ucsb.edu/