Replication presentation:

Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology

Martijn Schoonvelde, Gijs Schumacher, Bert N. Bakker

Introduction

Background:

Divide Between Political Science & Political Psychology in Text Analysis

- Different Approaches:
  - Political psychologists often use **supervised methods** (e.g., dictionaries) to analyze **stable characteristics** like personality type or linguistic styles.
  - Political scientists tend to use **unsupervised methods** (e.g., topic models, scaling models) to analyze **topical content or policy positions**.
- Lack of Cross-Disciplinary Work:
  - Collaboration between political scientists and political psychologists is rare ; both fields could benefit from **better integration**.

Introduction

Research question:

1. How do political scientists and political psychologists differ in their use of text analysis methods?
2. What are the implications of these methodological differences for interpreting political texts and behaviors?
3. How can integrating approaches from both disciplines improve the study of text in politics?

Four central steps for analysis:

Sampling texts (replication) ; Authorship as metadata ; **Preprocessing texts (replication)**; Analyzing texts (extension)

Methods

Sampling texts:

Dataset:

- **271 texts** from U.S. presidential and vice-presidential candidates:
  - **George W. Bush**, **John Kerry**, **Dick Cheney**, **John Edwards**.
- Sources:
  - **Network Interviews**
  - **Press Conferences**
  - **Town Hall Meetings**

Tool Used:

- **LIWC (Linguistic Inquiry and Word Count)** to measure linguistic dimensions.

Linguistic Dimensions Measured:

- **Cognitive Complexity**
- **Depression**
- **Honesty**
- **Presidentiality**
- **Aging**
- **Femininity**

Data Grouping & Filtering:

- Grouped by **speaker** and **text source**.
- Focused on candidates with **at least 10 speeches** (Bush and Kerry).

Methods

Sampling texts:

Analysis

- Mean & Standard Error Calculation:
  - Computed **mean LIWC scores** and **standard errors** for each linguistic dimension.

Visualization

- Bar Plot Creation:
  - Used *ggplot2* to create bar plots:
    - Displayed **mean standardized LIWC scores**.
    - Included **error bars** for variability.
    - Faceted by **speaker** and **text source**.

Results

Sampling texts:

Variation by Text Source:

- **Language use varies significantly** depending on the text source, even for the same speaker.

Influence of Context & Individual Differences:

- Both **political context** and **individual speaker traits** impact linguistic style.

Both political scientists and political psychologists have reason to believe that text sources are systematically different from each other so that
they should account for these differences in their models

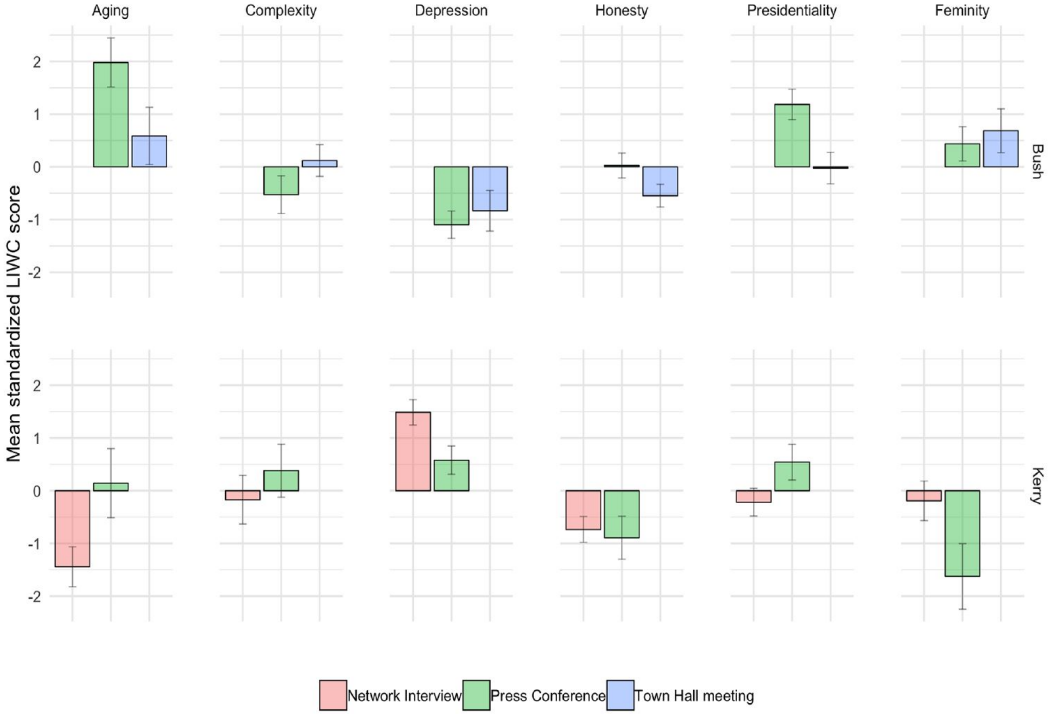


Figure 1. Linguistic style of George W. Bush and John Kerry on six linguistic style dimensions.

Note. This figure displays standardized LIWC scores for George W. Bush and John Kerry on six linguistic style dimensions (aging, complexity, depression, honesty, presidentiality, feminity) for various text sources for which we have at least 10 observations: network interviews (Kerry: n = 44), press conferences (Bush: n = 57: Kerry: n = 21), and town hall meetings (Bush: n = 38) (for more information, see Slatcher et al., 2007).

Methods

Pre-processing texts:

1. Data Collection and Preparation

- **Dataset:**
  - Loaded comb.corpus.RData containing speeches from European political leaders (3,301 speeches from 31 European leaders), with Left-Right (LR) and Progressive-Conservative (PC) ideological scores analyzed by the Comparative Manifesto Group database.
- **Data Cleaning:**
  - Removed irrelevant columns (e.g., eucount, eusentence).
  - Filtered out speeches with fewer than 200 tokens.
  - Limited data to speeches from the **European Parliament (EP)** and **National Leaders**

Methods

Pre-processing texts:

2. Text Preprocessing

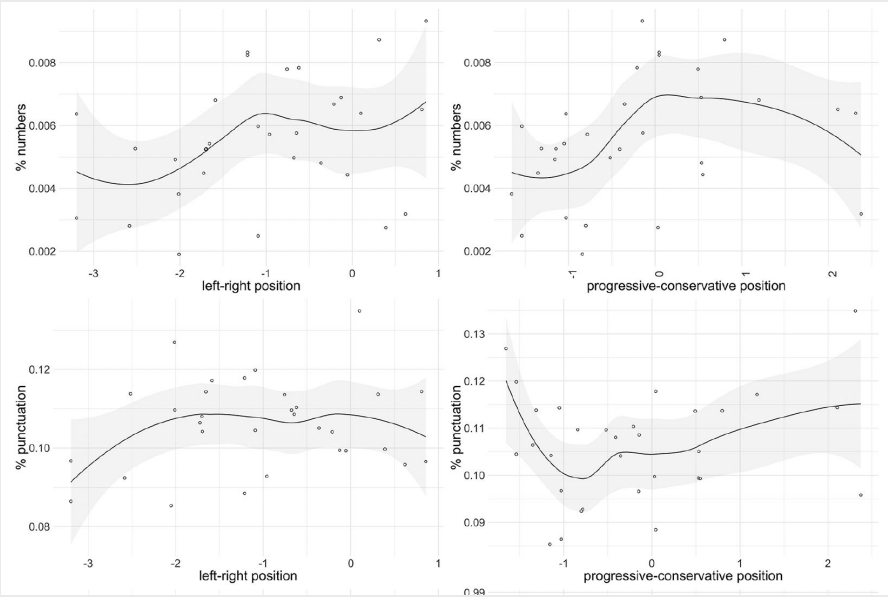
- **Tokenization and DFM Creation:**
  - Converted text into tokens using *quanteda*.
  - Created Document-Feature Matrices (DFMs) under different preprocessing conditions:
    - **Stopword Removal:** Excluded common English stopwords.
    - **Number Removal:** Removed numeric tokens.
    - **Punctuation Removal:** Removed punctuation marks.
    - **Stemming:** Applied word stemming to reduce words to their root forms.
- Calculated proportions of:
  - **Stopwords**
  - **Numbers**
  - **Stemmed Words**
  - **Punctuation**

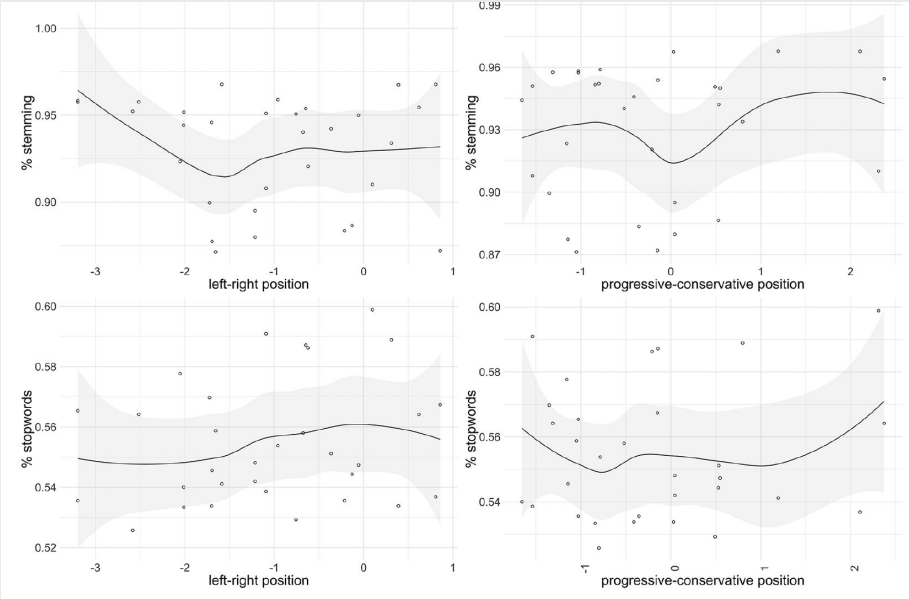
3. Visualization

- Plotted relationships between text features and ideological scores:
  - Correlation between text features and **Left-Right** positions.
  - Correlation between text features and **Progressive-Conservative** positions.

Results

Pre-processing texts:





Preprocessing steps like stopword removal, stemming, and punctuation removal can influence text analysis outcomes / These preprocessing steps are not indeed ideologically neutral

Autopsy

Pre-processing texts

Original:

```
#create no punctuation dfm
dtm.punctuation <- dfm(supercorpus, stem = FALSE, remove_punct = TRUE, remove_numbers = FALSE)
dtm.punctuation <- dfm_group(dtm.punctuation, groups = "speaker")
```

What we did:

```
#create punctuation as a proportion of text
# create dfm with the removal of puncuation
# tokenize and remove puncutation
tokens_no_punct <- tokens(supercorpus, remove_punct = TRUE)
# create dfm with
dtm_no_punct <- dfm(tokens_no_punct)
```

Extension

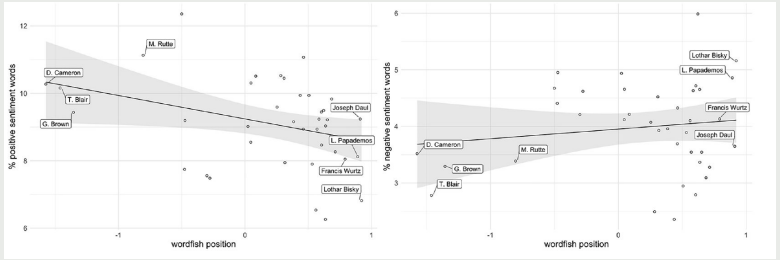
Extending Sentiment Analysis Using the EUSpeech Dataset

Original Findings (Section 4):

- Unsupervised learning model Wordfish model may inadvertently capture **sentiment** alongside **policy positions**.
- **Negative correlation** between Wordfish position and **positive sentiment words**.
- **Weak positive correlation** between Wordfish position and **negative sentiment words**.

Challenges:

- Co-occurrence models like Wordfish may be influenced by sentiment rather than policy.
- Example: *"Speaker A is more left-wing than Speaker B"* could be due to **higher sentiment word usage** rather than actual policy stance.
- This blurs the distinction between **emotional tone** and **policy position**.



Extension

Extending Sentiment Analysis Using the EUSpeech Dataset using different methods

Dataset:

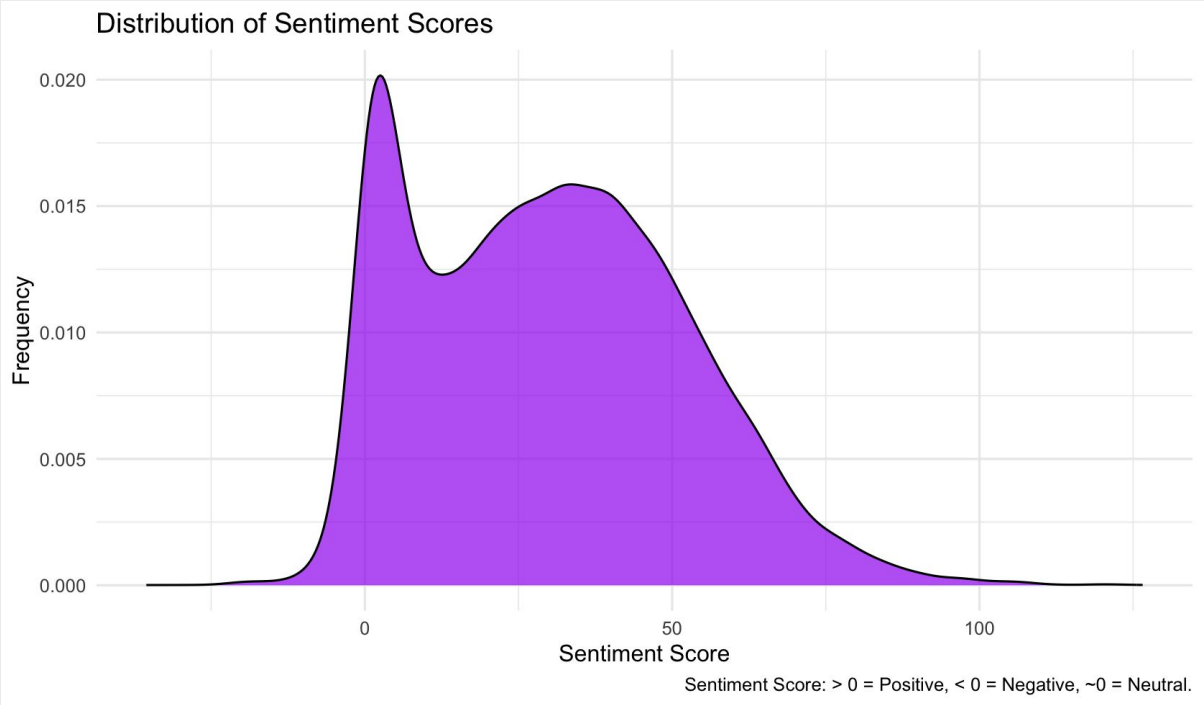
- Used the **EUSpeech dataset** containing speeches from European political figures.

Sentiment Scoring:

- Applied the **Syuzhet method** to calculate sentiment for each speech:
  - **Positive values** = Positive sentiment.
  - **Negative values** = Negative sentiment.
  - **Scores near zero** = Neutral tone.

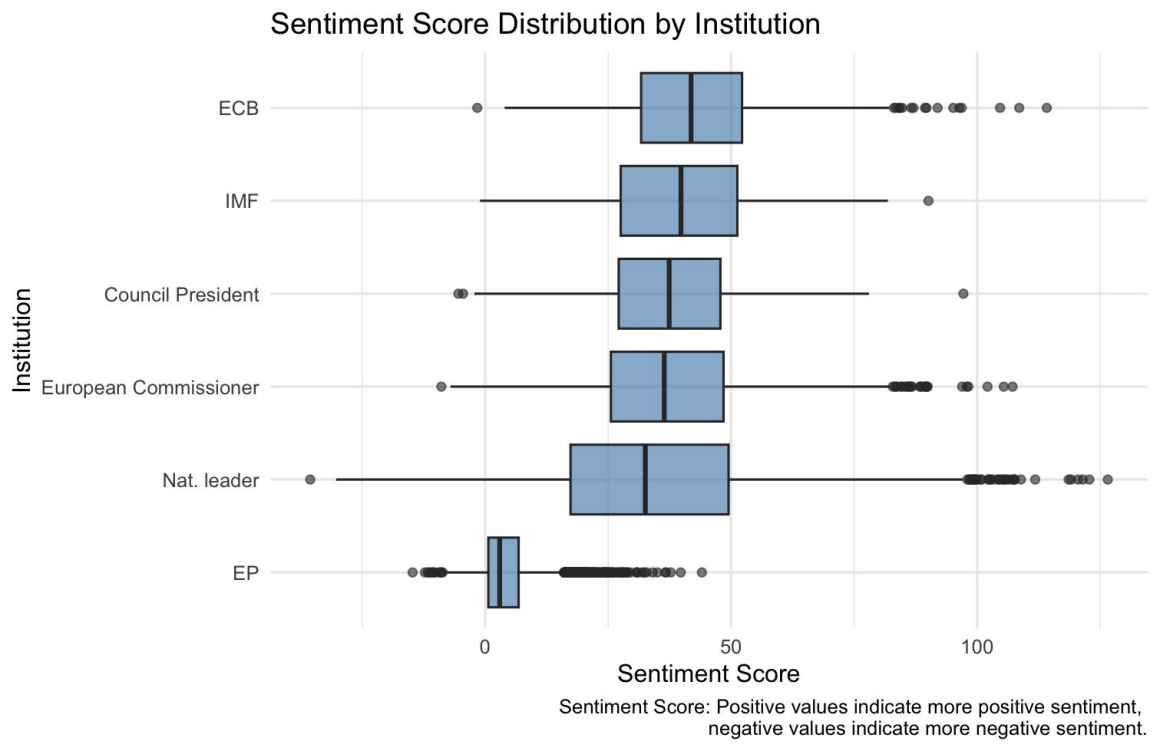
Extension-syuzhet

Figure 1: Histogram showing overall sentiment **distribution**



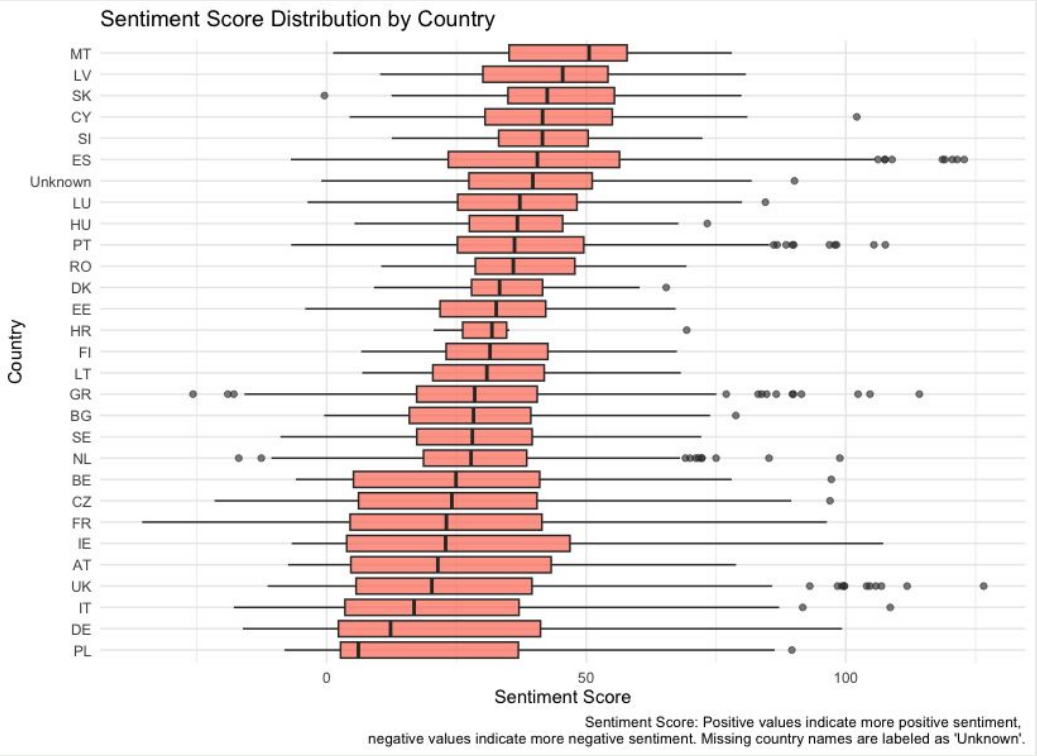
Extension-syuzhet

Figure 2: Boxplot of sentiment by **institution**



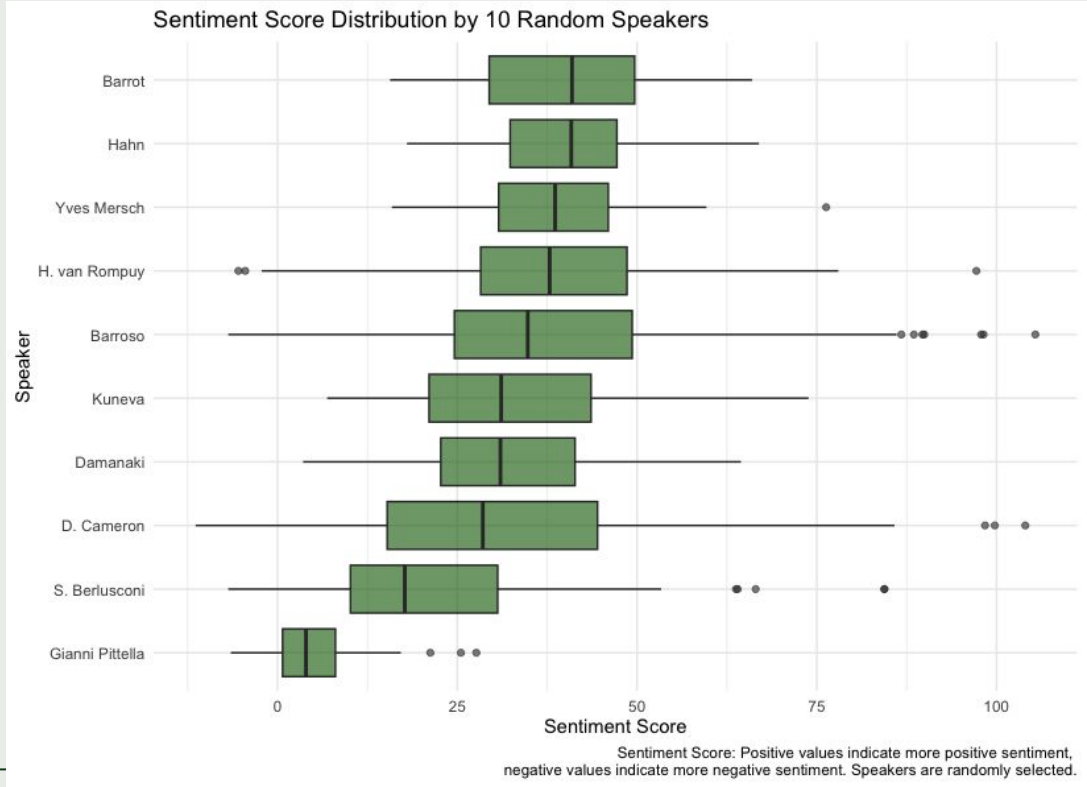
Extension-syuzhet

Figure 3: Boxplot of sentiment by **country**



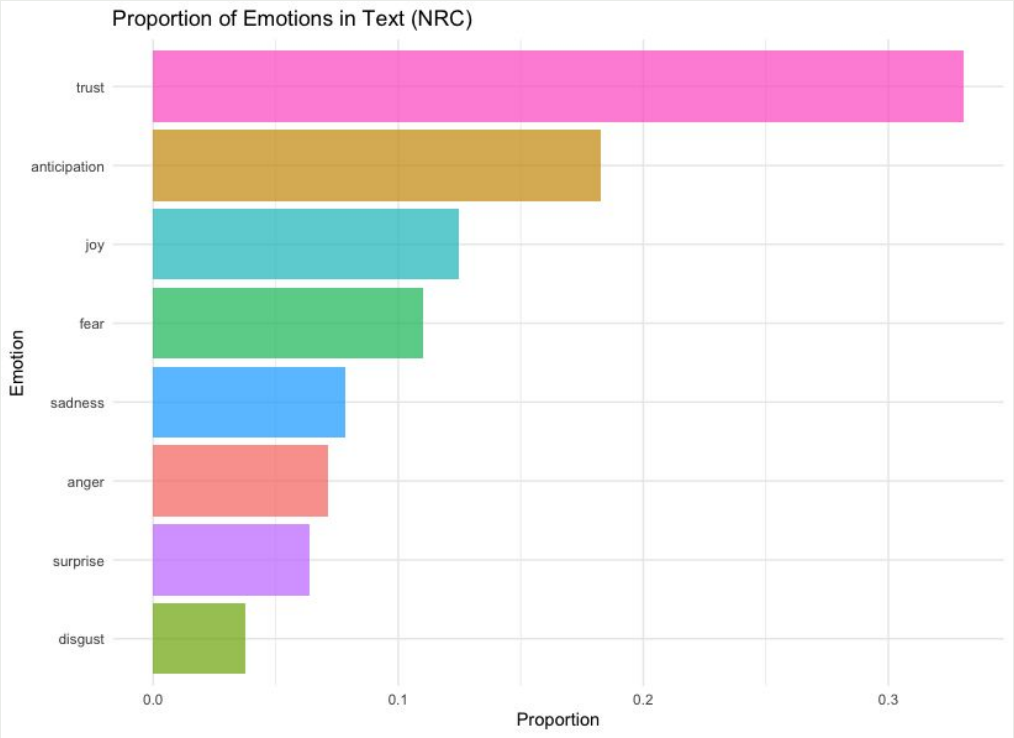
Extension-syuzhet

Figure 4: Boxplot of sentiment by **Speaker**



Extension-NRC

Figure 5: Overall Emotion Proportions



Extension-NRC

Figure 6: Emotion Words Proportions by **Institution**

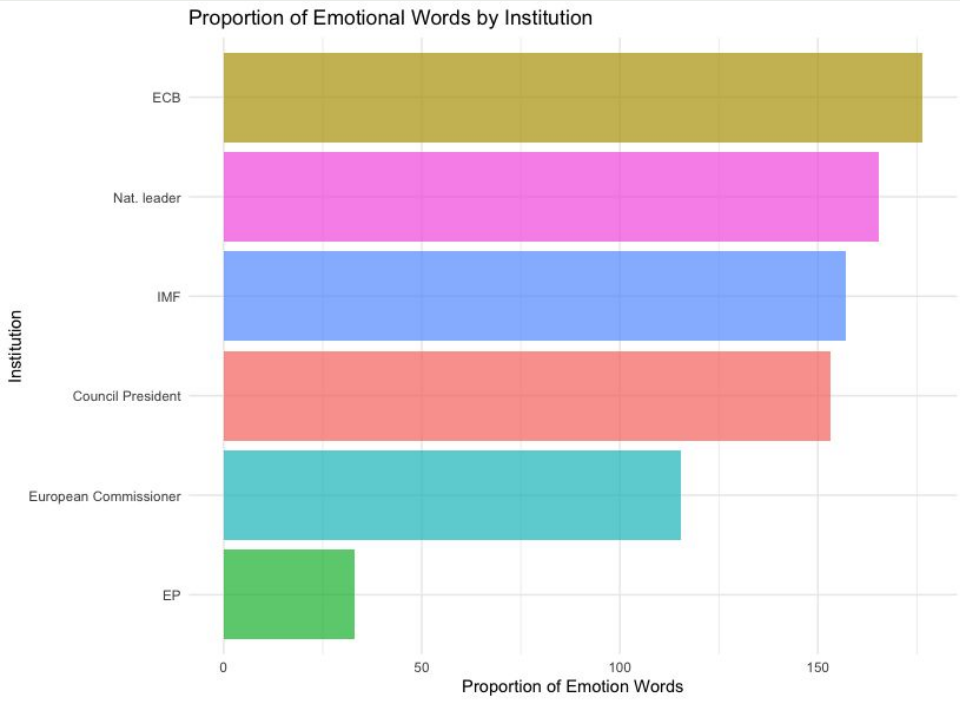
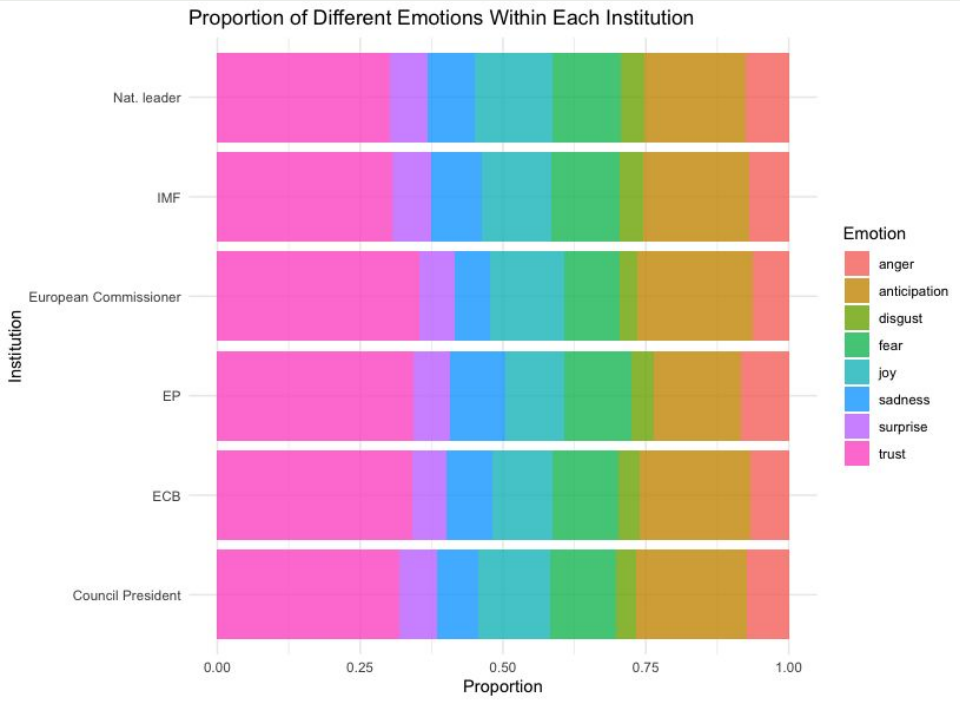
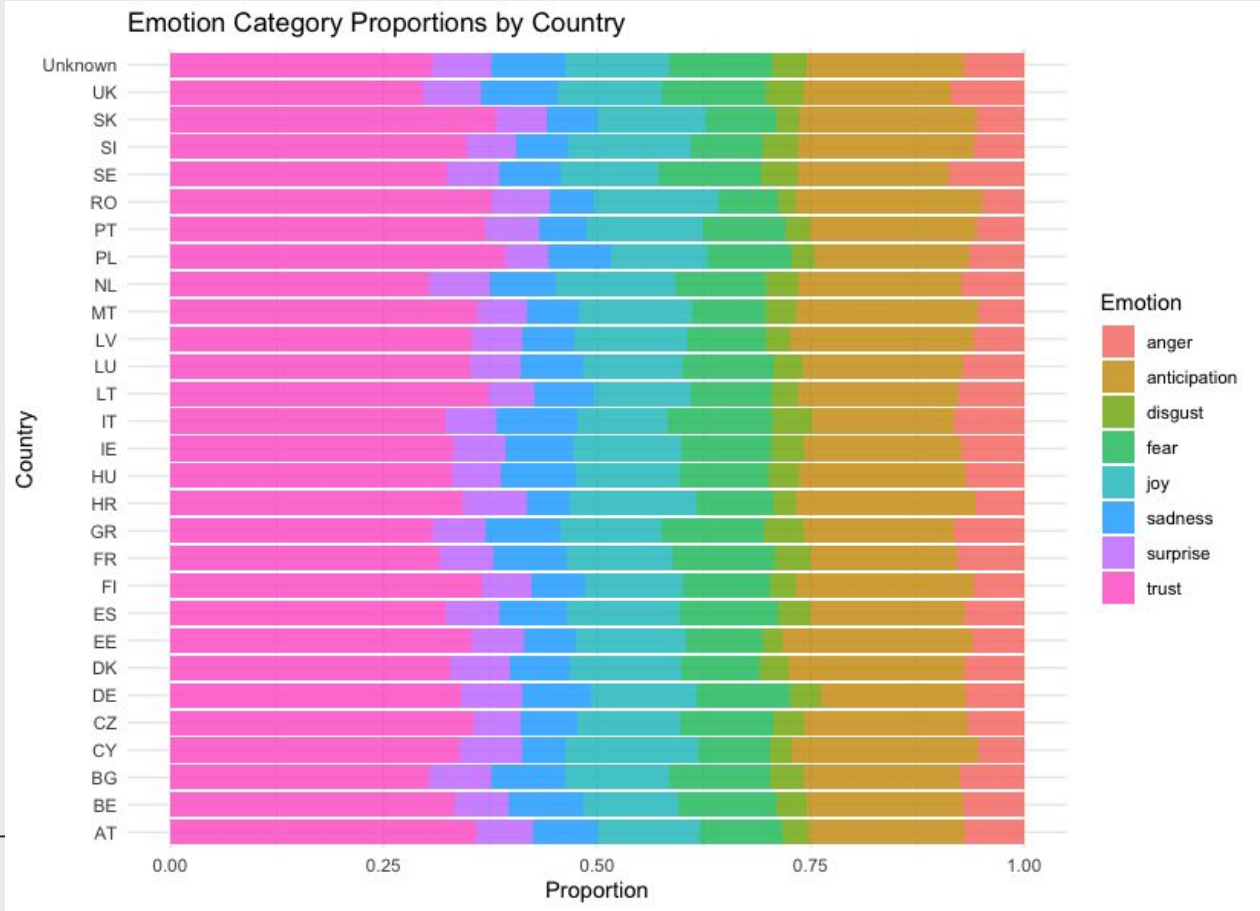


Figure 7: Different Emotion Proportions by **Institution**



Extension-NRC

Figure 8: Emotions Proportions by **Country**



Extension-NRC

Figure 9: Emotions Proportions by **Speaker**

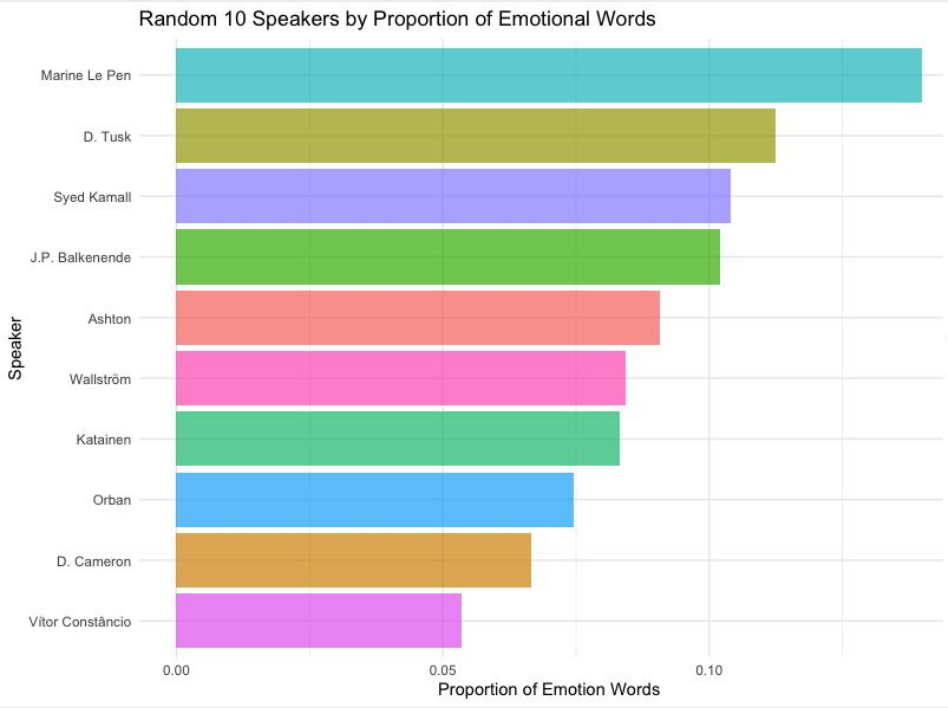
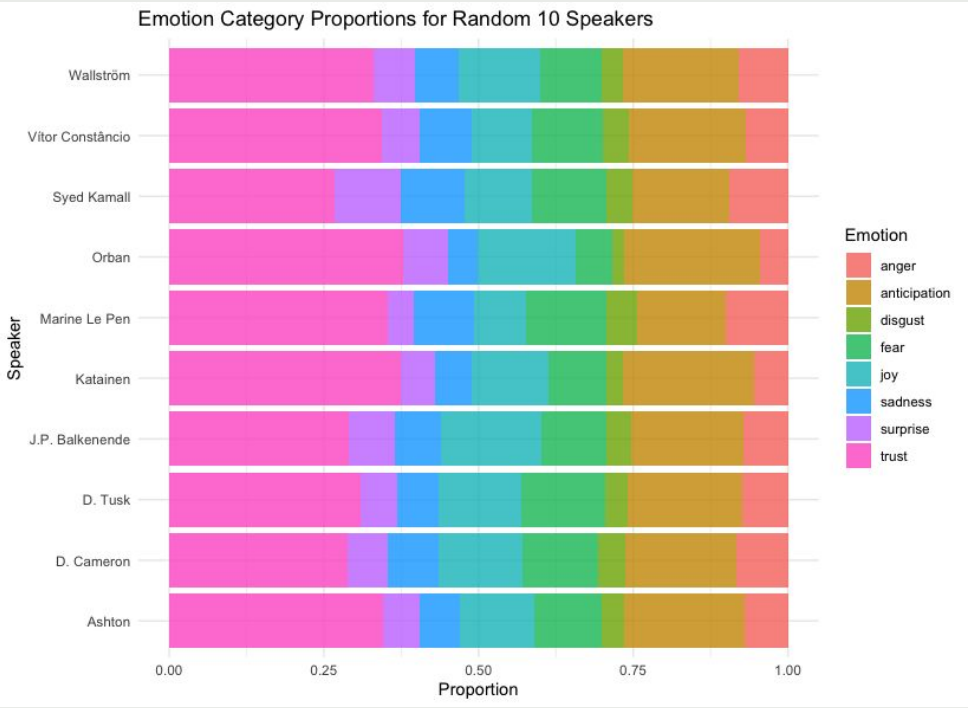


Figure 10: Emotions Category Proportions by **Speaker**



Extension

Extending Sentiment Analysis Using the EUSpeech Dataset using different methods

Sentiment Influence on Wordfish:

- Politicians using more **positive language** may appear to share similar policy positions in Wordfish models, even if their actual stances differ. The similarity in tone may just because of their personal speaking style.
- Politicians with different **proportions of emotional words** and **different combination of emotions** may also influence the policy positions in Wordfish's results

Key Takeaway:

- Political discourse in the dataset is characterized by **positive or neutral language using unsupervised learning models like Wordfish**, but this can **skew policy position estimates** if other considerations like sentiment are not properly accounted for.
- **Incorporating sentiment analysis** alongside Wordfish is essential to avoid misinterpreting **emotional tone** as **policy position**