Project Outline

Abraham Wang, Porter Stacy, Wentao Fan, Viggo Pauw

Note to CL: This is a first draft of the Project Outline document and includes two separate project outlines, a first choice and a back-up; this is due to the possibility of the first topic being too narrow or due to another group possibly choosing the same thing.

Furthermore, we did not include any notes for the Stand-Up Presentation as we feel this document touches upon the problem statement and its context, the data exploration and the additional data well enough already.

Project Outline: Toxic Political Speech and Polarization (First choice topic) (Working title: Polarized Words, Divided Society: Tracking Toxic Rhetoric in Dutch Politics)

1. Introduction

Political debates influence public opinion and government policies. However, in recent years, many debates have become more toxic and polarized, which may contribute to social inequality. Toxic speech can target certain groups, increase misinformation, and make political discussions less inclusive.

This project will analyze toxic speech in Dutch parliamentary debates and explore its connection with political polarization. We aim to answer the following **(research) questions**:

- On what political topics surrounding social inequality (like immigration and welfare) do debates contain the most toxic speech rhetoric? (Quantitative)
- Do certain political parties use toxic and polarising speech more often than others? (Quantitative)
- How toxic has polarising political speech in the Dutch Tweede Kamer become? And to what extent has the amount of polarisation increased over time? (Quantitative)
- Does political speech become more toxic during different times of the political season, for example before the national elections? (Quantitative)
- How does the general public perceive different kinds of political speech rhetoric regarding toxicity and polarisation? And how does this opinion differ when looking at their relevant political affiliation? (Qualitative)

- How does toxic political speech influence policy-making on social problems? (Qualitative)

Possible source to use for inspiration: https://horizonsproject.us/good-vs-toxic-polarization-twitt/

2. Data Sources

We will use the Tweedekamer 2024 dataset, which includes transcripts of Dutch parliamentary debates. The dataset contains full debates in tweede_2024.csv and two-minute debates in tweede_2024tweeminutendebat.csv.

Each dataset includes three key columns:

- filename: The name of the speech document.
- year: The year of the speech (2024).
- text: The speech content, for this datafile shortened to the first 10 lines. However, the full speech content is also available for use.

For our analysis, we will create our own dataset with all data we find relevant. For this, we will utilise the previously provided TK-analyser and TK-scraper GitHub-repositories. This final dataset will allow us to analyze how political parties discuss different topics and how toxic their relevant speech is.

For this topic idea, we assume to not have to use any additional datasets. However, if needed, we still see possibilities to add them to complement our data, for example by looking at polarisation trends in other countries or a dataset to compare our toxicity and polarisation scores and trends against.

3. Variables (DV, IV, Moderator, Mediator)

To structure our analysis, we (for now) defined the following variables:

The dependent variable (DV) is toxicity_score, which measures the toxicity level of each speech on a scale from 0 to 1, using NLP models.

Furthermore, we are thinking about adding a “polarisation score” metric to check for the severity of the polarisation. We, however, need to delve more into this in the future of the project to see if this is attainable and if so, how this metric could best be constructed.

The independent variables (IVs) include party (which represents the political party of the speaker), the speaker themselves, the topic (which identifies the policy area being discussed, also generated via ML-methods), speech type, which differentiates between full debates, two-minute debates, interruptions and any other form of debate in the dataset and the time-period of the speech.

The moderator variables include target_group, which examines whether the speech mentions marginalized groups, such as immigrants or low-income citizens.

The mediator variable is media_coverage, which checks whether the speech was widely reported in news and social media, as media attention may amplify the effects of toxic speech.

4. Project Structure Diagrams

For this part, we feel we are at a too early stage in the project to concretise this already. However, we are aware **qualitative and quantitative analyses**, **a database management structure** and a **data pipeline flowchart** need to be added here.

5. Method Analysis

For this part, we feel we are at a too early stage in the project to concretise this already. However, we are aware **model selection**, **the basic assumptions of the model**, **evaluation metrics** and **an assessment for the utility of the group project** need to be added here.

The project outline for the second topic starts on the next page

Project Outline: Political Speech in Debates vs. Political Speech in Online Profiling (Working title: Politics Unscripted: Debates vs. Digital Rhetoric)

1. Introduction

Political speech plays a crucial role in shaping public opinion and influencing policies. This political speech can be provided via multiple sources, of which debates in the Dutch Tweede Kamer (or parliament) are a good example. However, in the constantly evolving technological landscape of today, online political profiling has become a crucial part in this as well. The difference in these wildly different sources for political discourse can thereby sketch a picture of how these different sources are utilised to bring around political rhetoric and can thereby show the influence of these different sources. Also getting to grips with possible contradictions in political rhetoric here can be crucial to see how democracy in the modern world functions and how the political parties bring this towards the voter base.

This project aims to analyze speech patterns, stylistic differences and differences in political rhetoric in debates in the Dutch Tweede Kamer and videos on social media (most likely going to be YouTube), via answering the following **(research) questions**:

- To what extent is it possible to predict whether political rhetoric is taken from a debate or online political profiling? (Quantitative)
- How do the linguistic features and engagement metrics differ from Dutch debates and online political profiling? (Quantitative)
- How prevalent are contradictory statements from political parties between debates and online political profiling? (Quantitative)
- To what extent are there key differences in speech patterns and style between debate rhetoric and online political profiling? (Quantitative)
- How prevalent is reuploading (cut-up) debates to only show a particular part of the debate instead of the whole context? (The FvD-method) (Quantitative)
- To what extent is there a difference in speech rhetoric in debates towards videos that contain speech from academics and activists on the official YouTube channel of the political parties? (Quantitative)
- How does the general public view debate rhetoric compared to online political profiling? Does this differ between different demographics? (Qualitative)

2. Data Sources

We will use the Tweedekamer 2024 dataset, which includes transcripts of Dutch parliamentary debates. The dataset contains full debates in tweede_2024.csv and two-minute debates in tweede_2024tweeminutendebat.csv.

We will also incorporate data from YouTube, which provides political speeches, debates, and public discourse. By analyzing political speeches and debates uploaded on YouTube, we can examine speech patterns beyond official parliamentary records and compare them to the style, tone, toxicity levels and even the content of politicians' public speeches.

Each dataset includes three key columns:

- filename: The name of the speech document.
- year: The year of the speech.
- text: The speech content for this datafile shortened to the first 10 lines. However, the full speech content is also available for use.

For our analysis, we will create our own dataset with all data we find relevant. For this, we will utilise the previously provided TK-analyser and TK-scraper GitHub-repositories. This final dataset will allow us to analyze how political parties discuss different topics and how toxic their relevant speech is.

A comparable dataset will be created for the political YouTube videos, which will include the title of the YouTube video, the date of it being published and its text transcript.

Additional data could be political discourse on different social media platforms like Instagram, Facebook or Twitter/X (if X Æ A-12's father allows it).

3. Variables (DV, IV, Moderator, Mediator)

Dependent Variable (DV)

Speech category → Binary classification: Debates vs. Online Political Profiling

Independent Variables (IVs)

- The party of the person speaking.
- Time-period of the debate or video.
- The sentence structure and word choice
- Topic focus: Subjects discussed in the speech (policy, economics, activism, education).
- Speech context: Debate, formal address, rally, or informal remarks.

Moderator Variables

- Intended audience: Policymakers, general public, employees, activists, etc.

Mediator Variables

- Media coverage: Whether the speech received significant media attention, influencing its impact.

4. Project Structure Diagrams

For this part, we feel we are at a too early stage in the project to concretise this already. However, we are aware **qualitative and quantitative analyses**, **a database management structure** and a **data pipeline flowchart** need to be added here.

5. Method Analysis

For this part, we feel we are at a too early stage in the project to concretise this already. However, we are aware of model **selection**, **the basic assumptions of the model**, **evaluation metrics** and **an assessment for the utility of the group project** need to be added here.