

CMV2BAS Dataset

ChangeMyView Reddit Discussions to Bipolar Argument Structures Dataset



Funded by
the European Union



UK Research
and Innovation

Usage/License



Derived from the Webis ChangeMyView Corpus 2020 (<https://zenodo.org/records/3778298>). The original corpus is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Download



The dataset preparation pipeline and documentation are available at: <https://github.com/The-obsrvr/CMV-BAS-Data-Pipeline>

Researchers



- Siddharth Bhargava, FBK Povo & UDC A Coruña
- Sara Tonelli, FBK Povo
- Patricia Martín Rodilla, UDC A Coruña & IEGPS-CSIC Spain

Summary

The **CMV2BAS Dataset** is a curated collection of discussion threads derived from the public Webis ChangeMyView Corpus 2020 (CMV). Each thread represents a tree-structured, multi-party discussion categorized by whether the Original Poster (OP) changed their view (delta event). The dataset focuses on sensitive political and societal issues, including migration, European integration, climate change, and public health. It comprises 500 discussion threads annotated using POLIBRA, our LLM-assisted annotation framework. CMV2BAS supports research in argument mining, persuasiveness detection, stance detection, policy analysis, and political discourse analysis.

What We Offer

- An open-source, customizable data preparation pipeline for end-to-end conversion of discussion threads into Bipolar Argument Structures (BAS).
- A privacy-preserving, pseudonymized dataset for argumentative discourse research.
- A curated resource supporting research in political argumentation, social media analysis, persuasion, and long-form discourse analysis.

Key Features

- 500 discussion threads covering diverse political and societal topics: migration, climate policy, European integration, and public health.
- Annotated using POLIBRA, combining retrieval-augmented generation and self-consistency techniques for consistent long-form discourse analysis.
- Contains argumentative opinions connected through support and attack relationships.
- Annotation quality validated through manual human assessment on a representative subset.

Collaboration Objectives

- Promote collaboration between researchers in political science, sociology, social media studies, and argument mining.
- Facilitate study of long-form discourse and macro-level argumentation in domains such as political debates, multi-party discussions, legal reasoning, and scientific discourse.
- Support downstream research on persuasiveness detection, policy analysis, political decision-making, and argumentation strategies.

