DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates

Pierpaolo Goffredo, Deborah Dore, Elena Cabrio, Serena Villata

Université Côte d'Azur, CNRS, INRIA, I3S, France

{firstname.surname}@univ-cotedazur.fr

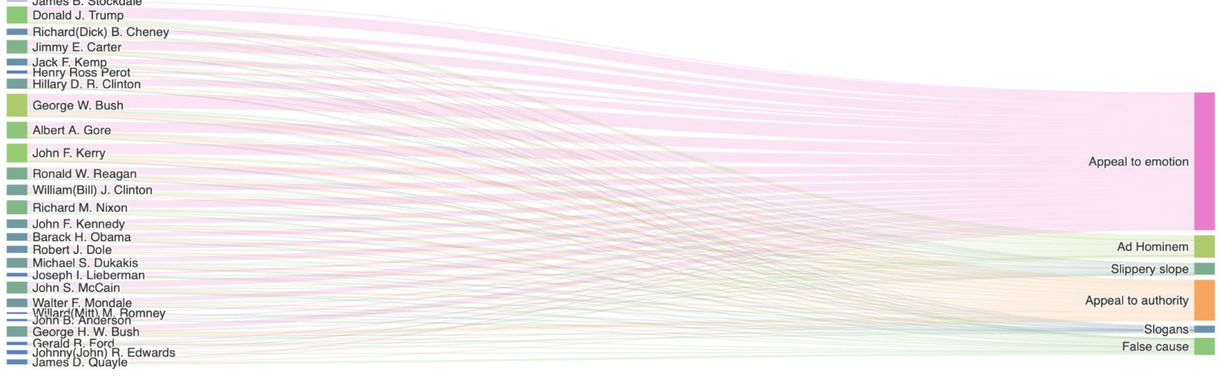
DISPUTool 3.0 At A Glance

DISPUTool is a web application designed for argumentation analysis of political debates that integrates state of the art methods to:

- **detect and classify arguments components** (*claim and premises)* ;
- **detect and classify relations** (*support, attack, equivalent);*
- **detect and classify fallacious arguments;**
- **rewrite fallacious arguments into a clearer and fairer version.**

Data Exploration

Users can explore the debates through multiple visualizations.



The ElecDeb60to20 Dataset

The **ElecDeb60to20** dataset contains all televised U.S. election presidential debates **from 1960 to 2020**. The dataset contains **44 debates** annotated with argument components, argument relations and fallacious arguments.

	Classes	Instances	Distribution
Argument Components	Claim	29624	53%
	Premise	26055	47%
	Total	*55679*	*100%*
Argument Relations	Attack	21687	85%
	Support	3835	15%
	Total	*25522*	*100%*
Fallacious Argument Components	Ad Hominem	341	12%
	Appeal to Emotion	1591	58%
	Appeal to Authority	433	16%
	False Cause	179	7%
	Slippery Slope	122	4%
	Slogans	78	3%
	Total	*2744*	*100%*

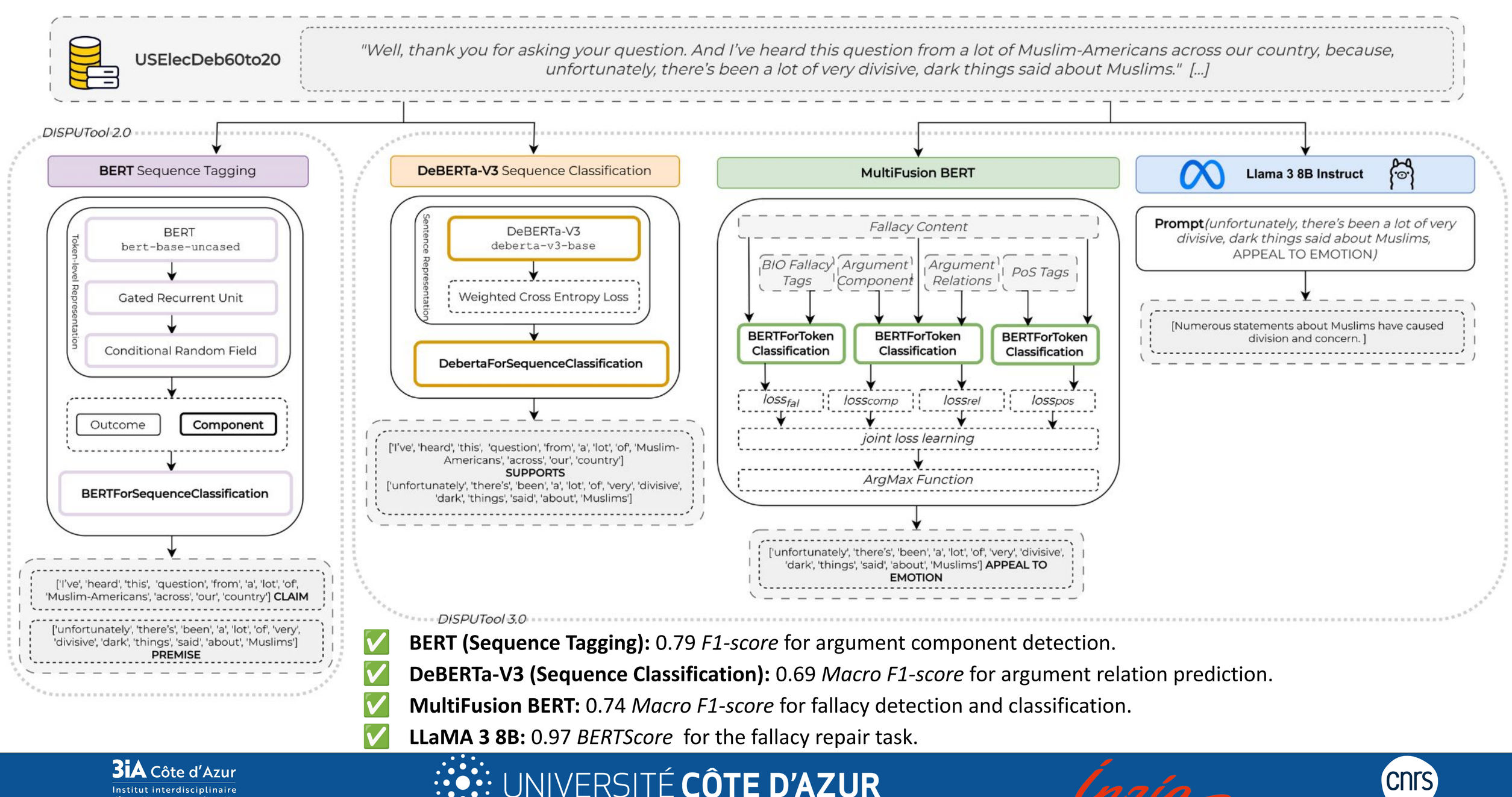
The FallacyFix Dataset

FallacyFix is a dataset comprising 747 manually repaired examples of fallacious arguments drawn from the ElecDeb60to20-fallacy dataset.

Subcategory	Frequency	Distribution
Loaded Language	416	56%
Flag waving	147	20%
Appeal to Pity	83	11%
Appeal to Fear	61	8%
Appeal to Popular Opinion	40	5%
Total	*747*	*100%*

FallacyFix is used to train and evaluate LLMs, such as LLaMA 3 8B and GPT-4, in **generating non-fallacious versions of fallacious arguments**. It supports different prompting strategies: *Zero-Shot, Few-Shot and Fine-Tuning.*

DISPUTool's Architecture



- ✅ **BERT (Sequence Tagging):** 0.79 *F1-score* for argument component detection.
- ✅ **DeBERTa-V3 (Sequence Classification):** 0.69 *Macro F1-score* for argument relation prediction.
- ✅ **MultiFusion BERT:** 0.74 *Macro F1-score* for fallacy detection and classification.
- ✅ **LLaMA 3 8B:** 0.97 *BERTScore* for the fallacy repair task.