Enhancing Discourse Parsing for Local Structures from Social Media with LLM-Generated Data-Researchers

Funded by the European Union

UK Research and Innovation

Source Datasets

The gold instances originate from subsets of three existing collections: the SFU Opinion and Comments Corpus (SOCC; Kolhatkar et al., 2020), the 2020 Tensions over Race and Heritage Collection (TRHC; Otero et al., 2021), and the Australian Election 2019 Tweets (AUSPOL). Users should refer to the original dataset licenses for terms and conditions.

Download

The dataset is available through the following GitHub repository: https://github.com/metabolean5/coling2025-jjes

Researchers

- Martial Pastor
- Nelleke Oostdijk
- Patricia Martín-Rodilla
- Javier Parapar

These researchers are affiliated with Radboud University nad Universidade da Coruña.

Summary

This dataset comprises 1,170 local RST (Rhetorical Structure Theory) discourse structures — including 900 LLM-generated synthetic examples and 270 gold-standard annotated examples — capturing the JOINT JOINT EVALUATION (JJE) coherence relation pattern across three social media platforms: online news comment sections (The Globe and Mail), a discussion forum (Reddit), and a social media messaging platform (Twitter). The dataset is designed to support research in discourse parsing for social media, with a focus on evaluative discourse structures prevalent in polarized discussions.

Key Features

- Multi-platform data: News comments, Reddit, and Twitter for cross-platform discourse analysis.
- Gold + synthetic annotations: 270 human-annotated cases plus 900 GPT-4-turbo (zero-shot) generated instances.
- RST-based labeling: Annotated with JOINT and EVALUATION relations in 4–6 EDU triadic structures.
- Research-focused: Supports discourse parser training and analysis of evaluative language in polarized discussions.

What We Offer

- Annotated RST structures: Gold-standard JJE instances with refined guidelines for social media (sarcasm, phatic expressions, etc.).
- Synthetic training data: 900 LLM-generated JJE structures in .mermaid format for parser training.
- Binary classification setup: JJE vs. non-JJE, using structurally similar JJX cases for realistic conditions.

Collaboration Objectives

- Improve RST discourse parsing on social media using LLM-generated synthetic data.
- Address the lack of RST-annotated resources for social media texts.
- Support research on evaluative discourse in polarized discussions.
- Promote synthetic data use for other RST constellations beyond JJE.

hybrids

HYBRIDS Website
www.hybridsproject.eu

Get in Touch
martial.pastor@gmail.com