CheckThat! 2025 - Task 1 on Subjectivity Detection: Instructions for Annotators

Introduction

This document is designed to instruct annotators on the task of subjectivity detection. In particular, the document contains information about data collection, the annotation process, the estimated workload, and the scheduled work plan.

Objective. Your objective as an annotator is to collect and annotate sentences from news articles to build a secret test set for the CLEF shared task on subjectivity detection.[1]

Annotation Deadline. The internal deadline for submitting the annotated dataset is set to the **17th of March 2025**.

Material All useful material is available at `https://github.com/nlp-unibo/checkthat--subjectivity-detection`.

Data Collection

What and from where do you have to collect data?

Number of samples. ∼300 sentences from about 3-4 articles.

[1] `https://checkthat.gitlab.io/clef2025/task1/`

Where to collect. News outlets covering controversial political affairs on law, civil rights, economics, and internal politics.

Language. *INSERT LANGUAGE*.

Instructions

In summary, the instructions for data collection are as follows:

1. Identify 3-4 news articles from outlets that cover controversial topics, suggesting an abundance of subjective content.
2. Download the news articles and split them into sentences.
3. Sample ∼300 sentences from download articles for annotation.

Support

If you have no experience in crawling news articles or programming, send us an email with the list of identified news articles and their URL (see Contact section). We will provide you with the crawled news articles for annotation.

Data Annotation

We provide detailed annotation guidelines for instructing you on the task.

Annotation Task. The annotation task is a binary annotation where a sentence is labeled as subjective (SUBJ) or objective (OBJ).

Guidelines. We provide English guidelines in our repository. Guidelines should apply to other languages with minor modifications (e.g., criteria-specific examples). The guidelines are short and schematic. **According to our estimates, it should take you no more than 15 minutes to read them carefully.**

Annotator Eligibility. We provide a small set of annotated examples, denoted as validation set, to validate your understanding of the provided annotation guidelines. **If you are new to the task**, we invite you to annotate these examples and compare your labels with those provided.

Number of Annotators. We suggest having **at least** three annotators per sample since it is a highly subjective task. Still, we prioritize annotations on previously uncovered languages rather than annotation robustness.

Estimated Workload. On average, annotating ∼100 sentences takes ∼1 hour for an unexperienced annotator and ∼30 minutes for an experienced one. **In total, we ask you no more than ∼2 hours of your time over 2 months.**

Annotation Methodology.

We recommend the following annotation methodology, if possible. First, each annotator labels collected sentences, and inter-annotator agreement (IAA) is computed. Second, identify disagreement cases, discuss them to reach an agreement, if possible, and re-compute IAA. If the number of annotators is even (e.g., 2), consider instructing a third annotator to break disagreements. If there are several annotators, consider splitting sentences into groups to reduce the annotators' workload. Note that some sentences should be shared among all annotators to assess IAA robustly. See [Antici et al., 2024] (Section 2) for more details about the annotation methodology.

Instructions

In summary, the instructions for data annotation are as follows:

1. Read annotation guidelines carefully.
2. (Optional) Validate your understanding by annotating the provided validation set.
3. Follow the provided annotation methodology to label collected sentences.
4. Prepare your annotations in a .csv file with header **sentence, label**.

5. Send the .csv file via email (see **Contact**).

Contact

Please keep every contact person in cc when sending an email. Feel free to contact us for any issues or doubts!

- (Federico Ruggeri): federico.ruggeri6@unibo.it
- (Katerina Korre): aikaterini.korre2@unibo.it
- (Arianna Muti) arianna.muti@unibocconi.it

References

[Antici et al., 2024] Antici, F., Ruggeri, F., Galassi, A., Korre, K., Muti, A., Bardi, A., Fedotova, A., and Barrón-Cedeño, A. (2024). A corpus for sentence-level subjectivity detection on English news articles. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 273–285, Torino, Italia. ELRA and ICCL.