



Updated Oct 22



Share

[NL] OIE Data Collection

By William Hsu 2 min 2 Add a reaction



Open IE Task 1: Data Collection from Social Platforms

Team Members

 **Team Lead:** [James Chapman](#)Preview

- **Other Team Members:** [Jahid Choton](#)Preview, [Chipo Sekabanja](#)Preview

Overview

 material to import from task thread of Teams channel

Prompt [Do those four steps in a single stepwise... | Try in Copilot Chat](#)

Do those four steps in a single stepwise iteration for each prompt, aggregating your results and condensing your reply for conciseness.

A brief synopsis of Task 1, accessible to non-specialists in the technical AI, ML, NL, CV, and HCI/HRI areas canvassed above, is:

Data collection from social platforms

This task gathers publicly available posts, videos, and podcasts from platforms like YouTube, Reddit, and X/Twitter that relate to a scientific or ethical topic. Using text and multimedia analysis, it translates speech, captions, and visual text into searchable text fields. The goal is to capture diverse public perspectives, examples, and informal reasoning. These posts form a living dataset for downstream AI learning. By organizing this content, we enable deeper study of how people discuss complex issues—helping AI systems learn from real-world dialogue rather than static documents.

Response

Great! Here's a **stepwise literature review and planning iteration** for **Task 1: Data Collection from Social Platforms**,

based on your synopsis and goals.

❖❖ Task 1: Literature Review Summary

1. Relevant Papers & Methods

From SIGIR, KDD, WSDM, and AAAI:

- **Efficient Log Parsing with LLMs** (KDD 2024): Uses transformer-based models to extract structured data from noisy logs — applicable to social media scraping.
- **LabelCraft** (WSDM 2024): Automates labeling of short videos using metadata and user behavior — relevant for crawler prioritization.
- **GoldCoin** (IJCAI 2024): Contextual integrity for privacy-aware data collection — important for ethical scraping.
- **Structured Prompts for Debiasing LMs** (AAAI 2024): Can guide crawler prompts to avoid biased data collection.

2. Key Contributions & Gaps

• Contributions:

- Use of LLMs for parsing and labeling unstructured content.
- Ethical frameworks for privacy-aware scraping.
- Adaptive crawling based on user behavior and content features.

• Gaps:

- Limited multimodal integration (video, audio, OCR).
- Few systems address **living datasets** or **continuous updates**.
- Sparse work on **scientific discourse** in social media.

3. Experimental Design Suggestions

- **Baseline:** Compare static scraping vs. smart crawler with feedback loop.
- **Metrics:** Coverage (topic diversity), freshness (update rate), relevance (scientific/ethical alignment).
- **Tools:** Use ASR + OCR pipelines for video/audio posts; embed metadata for downstream tasks.
- **Ethics:** Integrate contextual integrity checks (GoldCoin-style) for privacy compliance.

4. Recommended Venues & Contingency

- **Primary:** SIGIR, WSDM, The Web Conference (systems-oriented)
- **Secondary:** UMAP (personalization), ICWSM (social media), ASONAM (network modeling)
- **Contingency:** Workshops at SIGIR or WSDM; IUI for interactive systems; EAAMO for ethical scraping

Next task - Task 2: Learning to Rank

Products

Scraping API

LLM prompts

- [GPT-5] [Query Generation](#) - c. 30 Sep 2025

Index

Research Planning

- [NL] [OIE Data Collection](#) | a) [Subtasks](#)Preview
- [NL] [OIE Data Collection](#) | b) [State of the Art](#)Preview
- [NL] [OIE Data Collection](#) | c) [Implementations](#)Preview
- [NL] [OIE Data Collection](#) | d) [Benchmarks](#)Preview
- [NL] [OIE Data Collection](#) | e) [Application Domains](#)Preview
- [NL] [OIE Data Collection](#) | f) [Datasets](#)Preview
- [NL] [OIE Data Collection](#) | g) [Experiments](#)Preview

Literature Review

- a) [NL] [OIE Data Collection](#) | Literature Review: [Subtasks](#)Preview
- b) [NL] [OIE Data Collection](#) | Literature Review: [SOTA](#)Preview
- c) [NL] [OIE Data Collection](#) | Literature Review: [Implementations](#)Preview
- d) [NL] [OIE Data Collection](#) | Literature Review: [Benchmarks](#)Preview
- e) [NL] [OIE Data Collection](#) | Literature Review: [Application Domains](#)Preview
- f) [NL] [OIE Data Collection](#) | Literature Review: [Datasets](#)Preview
- g) [NL] [OIE Data Collection](#) | Literature Review: [Experiments](#)Preview

Development

- a) [NL] [OIE Data Collection](#) | Development: [Subtasks](#)Preview
- b) [NL] [OIE Data Collection](#) | Development: [SOTA](#)Preview
- c) [NL] [OIE Data Collection](#) | Development: [Implementations](#)Preview
- d) [NL] [OIE Data Collection](#) | Development: [Benchmarks](#)Preview
- e) [NL] [OIE Data Collection](#) | Development: [Application Domains](#)Preview
- f) [NL] [OIE Data Collection](#) | Development: [Datasets](#)Preview
- g) [NL] [OIE Data Collection](#) | Development: [Experiments](#)Preview

Results

- a) [NL] OIE Data Collection | Results: SubtasksPreview
- b) [NL] OIE Data Collection | Results: SOTAPreview
- c) [NL] OIE Data Collection | Results: ImplementationsPreview
- d) [NL] OIE Data Collection | Results: BenchmarksPreview
- e) [NL] OIE Data Collection | Results: Application DomainsPreview
- f) [NL] OIE Data Collection | Results: DatasetsPreview
- g) [NL] OIE Data Collection | Results: ExperimentsPreview

a) Subtasks

Literature Review: Subtasks



Development: Subtasks

Results: Subtasks

b) State-of-the-Art

Literature Review: SOTA



Development: SOTA

Results: SOTA

c) Implementations

Literature Review: Implementations



Development: Implementations

Results: Implementations

d) Benchmarks

Literature Review: Benchmarks



Development: Benchmarks

Results: Benchmarks

e) Application Domains

Literature Review: Application Domains



Development: Application Domains

Results: Application Domains

f) Datasets

Literature Review: Datasets



Development: Datasets

Results: Datasets

g) Experiments

Literature Review: Experiments



Development: Experiments

Results: Experiments

Navigation

- [\[NL\] Chat](#) | [Channel](#) | [Team](#) | [Wiki](#)
- [\[OIE\] Chat](#) | [Channel](#) | [Plan](#) | [Lit](#) | [Dev](#) | [Wiki](#)

Return

- Open IE Tasks Page - [\[NL\] Open IE Tasks](#)
- Open IE Project Home Page - [\[NL\] Open Information Extraction \(OIE\)](#)
- KDD Lab Natural Language (NL) Division Page - [Natural Language](#)
- KDD Lab Research Divisions Page - [Divisions](#)
- KDD Lab Research Page - [Research](#)
- **KDD Lab Home Page** - [!\[\]\(2c0c58cb268bb8420b6fc93187c8f293_img.jpg\) KDD Lab](#)

Add a comment

Add a reaction