Enriching EUDAMED with Web Data: A Semi-Supervised Multi-View NLP Approach

Riccardo Gibello[1] and Enrico Gianluca Caiani *[1,2]

[1]Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy
[2]IRCCS Istituto Auxologico Italiano, San Luca Hospital, Milan, 20149, Italy

February 14, 2026

1 Context

The European Database on Medical Devices (EUDAMED) collects structured information about medical devices in the EU, such as UDI codes and manufacturer details. However, it often lacks detailed free-text descriptions of device functionality. This limits more advanced analysis and automated monitoring. This project aims to complement EUDAMED by collecting and processing unstructured data from publicly available web sources.

2 Objective

The goal is to create an open dataset of medical device descriptions from web data. You will develop a complete Machine Learning pipeline to process, filter, and extract useful information to enrich EUDAMED.

The main tasks are:

1. **Data Collection & Engineering:** Build and maintain a web crawler to gather structured metadata and free-text content from manufacturer websites.
2. **Gold Standard Creation:** Annotate data using tools like `Label Studio` [1, 8] to produce high-quality ground truth for:
   - *Page Classification:* Identifying valid medical device pages.
   - *Named Entity Recognition (NER):* Extracting entities like device names, therapeutic areas, and intended uses [6].
3. **NLP Pipeline Development:** Build an ML pipeline to clean data and extract knowledge:
   - Train a classifier to filter relevant web pages.
   - Train NER models to map unstructured text to structured EUDAMED fields.

*Advisor

3 Expected Skills

By completing this project, the student will gain practical experience in:

- **End-to-End ML Workflow:** From data collection and cleaning to annotation, model training, and evaluation.
- **NLP Techniques:** Text classification and entity extraction using modern libraries.
- **Data Curation:** Creating reliable annotated datasets, an important skill in ML projects.
- **Evaluation:** Understanding precision, recall, F1-score, and debugging models on noisy web data.

4 Research Aspect: Semi-Supervised Learning

As an extension of the engineering-focused project, it is possible to explore **semi-supervised learning** to improve model performance when labeled data is scarce. A *Multi-View Learning* approach [2–5, 7, 9] could be applied, leveraging multiple complementary views of web pages, such as:

- **Page text:** The raw content of the page.
- **HTML metadata and tags:** Structured elements that provide semantic cues.
- **URL structure:** Patterns that may indicate device type or manufacturer.

The main idea is to train separate classifiers on each view and use confident predictions from one to augment the training data of the other(s), in a **co-training or tri-training framework**. Iteratively, this can expand the labeled dataset and improve performance.

Possible challenges include:

- Ensuring that the views provide sufficiently complementary information.
- Preventing the propagation of incorrect pseudo-labels between classifiers.
- Balancing model complexity with limited labeled data.
- Adapting classical co-training ideas to modern text representations (e.g., transformer embeddings).

Exploring this direction would offer insights into how web-derived multi-view features can be combined effectively in a semi-supervised NLP pipeline.

5 Candidate Profile

We are looking for students who are motivated to experiment and learn.

- **Background:** Computer Science, Biomedical Engineering, or related fields.
- **Programming:** Proficiency in Python. Experience with web scraping (Scrapy, BeautifulSoup) or data processing (Pandas) is a plus.

- **ML/NLP Interest:** Basic understanding of ML concepts. Curiosity and willingness to learn libraries like `PyTorch`, `HuggingFace`, or `TensorFlow`.
- **Scientific Mindset:** Interest in semi-supervised learning and producing well-documented, reproducible results.

References

[1] Label Studio OSS: Open source data labeling platform. `https://labelstud.io/label-studio-oss/`, 2026. Accessed: 2026-02-13.

[2] Kamal Berahmand, Fatemeh Daneshfar, Maryam Rahmaninia, Maryam Haghighat, and Mahdi Jalili. A comprehensive survey on multi-view classification: Methods, applications, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–34, 2025.

[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[4] Yitao Ding, Mohamed Shalaby, Narinderjit Singh Sawaran Singh, and Raed HC Alfilh. Multi-view text classification through integrated rnn autoencoder learning of word, sentence, emotion and paragraph representations. *Scientific Reports*, 2025.

[5] Hongyu Guo, Colin Cherry, and Jiang Su. End-to-end multi-view networks for text classification, 2017.

[6] Jinhui Li, Aixin Sun, Jialong Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022.

[7] Md Mezbaur Rahman and Cornelia Caragea. Llm-guided co-training for text classification, 2025.

[8] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. Open source software available from https://github.com/HumanSignal/label-studio.

[9] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.