Enhancing Manufacturing Data Accessibility Through Web Scraping and AI-Driven Entity Recognition

PURDUE UNIVERSITY | The Data Mine

Nikhil Venkatachalam, Dhruv Soni, Prateet Saluja, Akash Ravandhu, Harini Muthu, Shreya Kamath, Avnish Kanungo, Avi Aggarwal, Joann Ncube, Aidan McDonough, David Kim, Yi-Fang Cheng, Jingying Hu, Lynn Nakamura, Aaron Lee, Golsa Khodadadi

INTRODUCTION/BACKGROUND

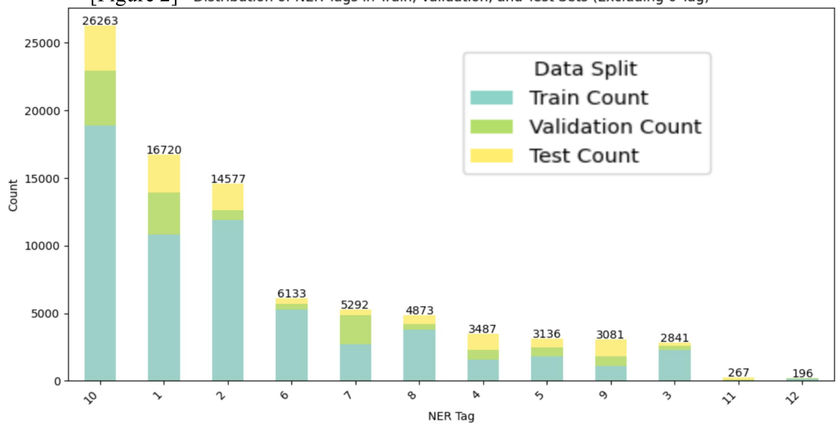
KNUDSEN INSTITUTE
for Applied Research on Surge Theory

The Knudsen Institute is a 501(c)(3) non-profit organization which works to advance America's defense industrial base surge capacity, with a focus on small & medium manufacturers. This project's focus is to explore the tools and methods for structured manufacturing-related data extraction.

RESULTS

- Web scraping: Developed and tested web scraping algorithms on 20+ manufacturer websites, scraping 1500+ sub-pages to gather relevant website data.
- Model training: Our current full fine-tuned model achieved precision 0.821, recall 0.822, and F1 score 0.821, with 0.935 accuracy.
- Data cleaning: The two inaccurate labels in the FabNER dataset "as" and "be" were detected (shown in Figure 1) and both corrected to 0 (non-category) during data cleaning.

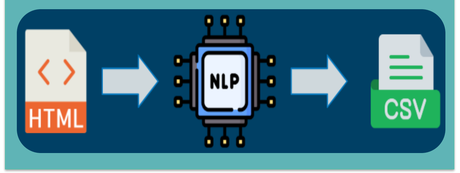
[Figure 2] Distribution of NER Tags in Train, Validation, and Test Sets (Excluding 0 Tag)



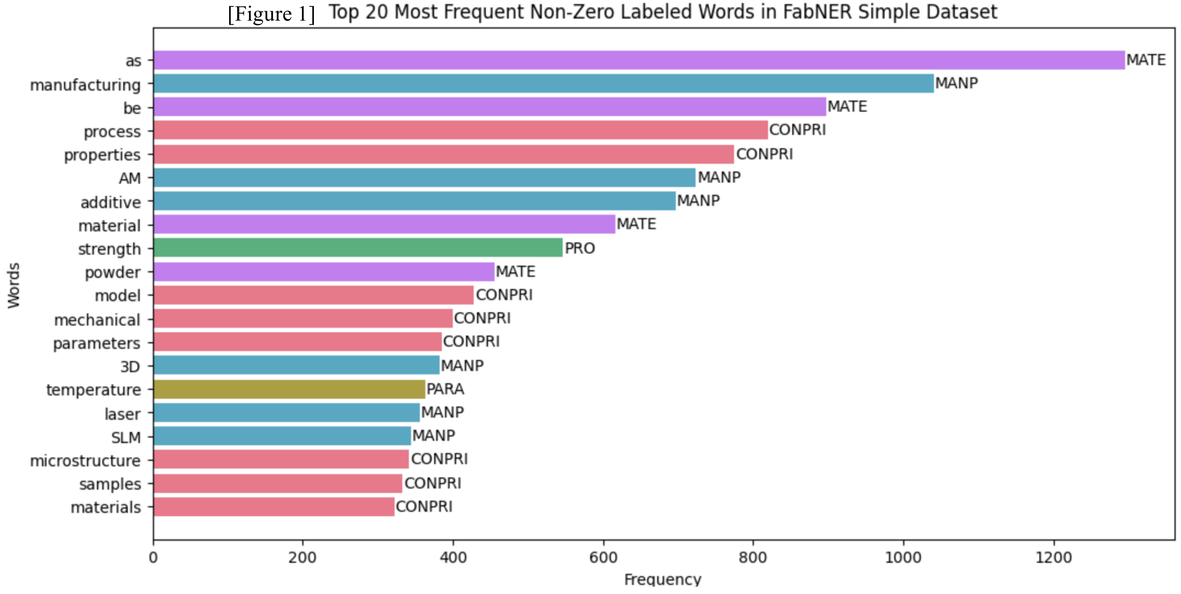
Research & Methodology

FIRST SEMESTER

We initially focused on developing web scraping techniques for manufacturing websites using BeautifulSoup and Selenium. These tools were used due to their efficiency in scraping static HTML as well as interacting dynamically with websites that require user interaction. We later experimented with using Large Language Models to convert unstructured web data into more useful structured data.



[Figure 1] Top 20 Most Frequent Non-Zero Labeled Words in FabNER Simple Dataset



SECOND SEMESTER

We worked on training a Named Entity Recognition (NER) model that can label words in a statement into different manufacturing categories. We have attempted methods including full fine-tuning, partial fine-tuning, and optimizing hyperparameters. We later learned that there were inaccuracies in the training data that we were using, so we tried manually correcting the tags of certain words throughout the dataset to make the data more accurate.

CONCLUSION

By combining web scraping with specialized NER models, we unify critical data for faster and accurate resource allocation, creating a supply chain management system for manufacturing SMEs (small and medium-sized enterprises). Our approach ensures scalable, high-precision entity detection that directly drives better resource planning and operational agility.

FUTURE STEPS

- Use our active learning pipeline on the testing dataset to help improve the NER model precision further
- Using the NER model on the data that has been obtained from the web scraping for creating the manufacturing specific database.
- Aggregating structured manufacturer data and compiling it into an evergreen dataset.

ACKNOWLEDGEMENTS

We'd like to sincerely thank The Knudsen Institute and our mentors: Richard Leu, Bhairavsingh Ghorpade, and Kevin Alexander. We also greatly appreciate the support of our TA, Simrith Ranjan, and The Data Mine Corporate Partners staff!