SYNCLAB : INTELLIGENT IMAGE DATA GENERATION

Ayush Raj and Priyansh Bhakuni | Prof. P. Senthilnathan | SCOPE

Introduction

Creating diverse test data for vision AI is challenging. Traditional methods involve manual website scanning, prone to biases. Existing datasets often lack specificity, restrictions on number of images and might face copyright issues. Our project proposes an NLP-driven solution to revolutionize dataset creation, particularly for emotion detection.

Motivation

Inspired by the challenges above, we aim to develop a solution streamlining laborious web scraping tasks for diverse vision AI test data. Synclab aims to revolutionizes web scraping, addressing accuracy and dataset limitations.

SCOPE of the Project

1. **Query Expansion:** Enhancing search queries for accuracy.

2. **Image Scraping:** Automated methods for data retrieval.

3. **Model Assessment:** Evaluating dataset quality using complex models.

4. **Data Augmentation:** Enhancing dataset diversity through augmentation.

Methodology

Addressing the challenge of laborious and time-consuming creation of test data for various vision AI solutions, this project aims to streamline the process through innovative methodologies. The primary focus is on enhancing efficiency and effectiveness in web scraping and image extraction, while also optimizing dataset quality for diverse use cases.

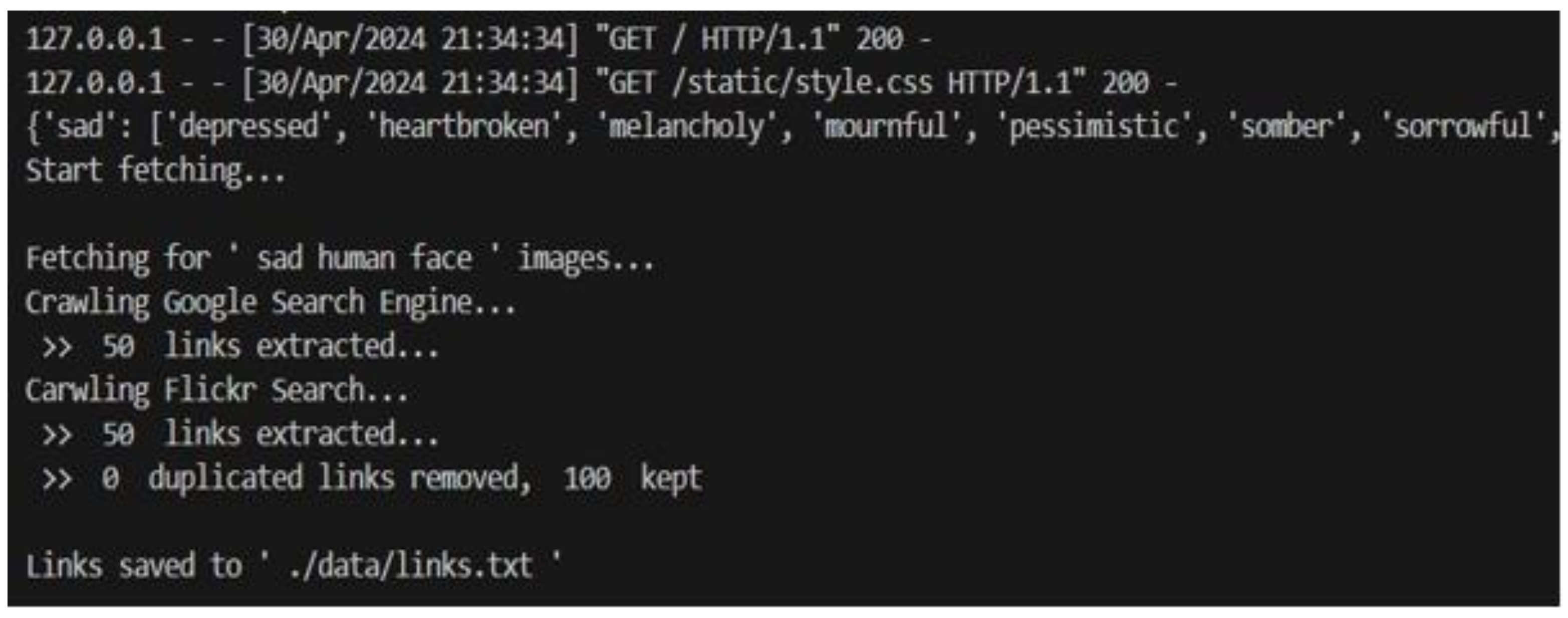
I. **Query Expansion Techniques:** Leveraging NLP models, we expand queries to retrieve a broader spectrum of keywords, facilitating comprehensive internet image searches and enhancing dataset richness.

II. **Image Scraper Methods:** Employing automated solutions via Selenium, Google and Flicker API we automate image dataset creation, ensuring adherence to specific requirements, thus eliminating manual labour and enhancing efficiency.

III. **Complex Model Assessment for Dataset Validation:** Employing a sophisticated Emotion Detection model founded on Complex Computer Vision and AI, we rigorously evaluate dataset quality, emphasizing precision and diversity for optimal model performance.

IV. **Data Augmentation:** Implementing image augmentation techniques also known as image post processing such as rotation, flipping, and scaling, we diversify and enrich the quality of the dataset, improving model robustness and performance.

This methodology ensures efficient, accurate, and diverse creation of test data for vision AI solutions, addressing the challenges posed by manual labor and generic search engines.



```
127.0.0.1 - - [30/Apr/2024 21:34:34] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 21:34:34] "GET /static/style.css HTTP/1.1" 200 -
{'sad': ['depressed', 'heartbroken', 'melancholy', 'mournful', 'pessimistic', 'somber', 'sorrowful',
Start fetching...

Fetching for ' sad human face ' images...
Crawling Google Search Engine...
 >>  50  links extracted...
Carwling Flickr Search...
 >>  50  links extracted...
 >>  0  duplicated links removed,  100  kept

Links saved to ' ./data/links.txt '
```

Figure 1 : Query Expansion for Web Scrapping (in the backend)

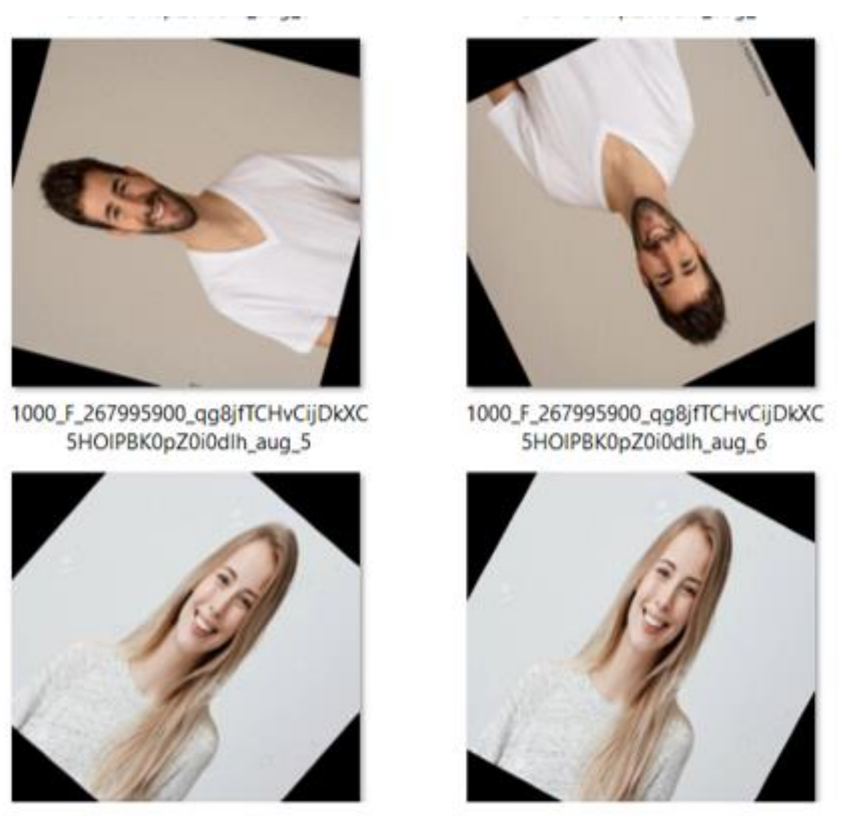

1000_F_267995900_qg8jfTCHvCijDkXC 5HOIPBK0pZ0i0dlh_aug_5
1000_F_267995900_qg8jfTCHvCijDkXC 5HOIPBK0pZ0i0dlh_aug_6

Figure 2 : Data Augmentation for Enriched Dataset

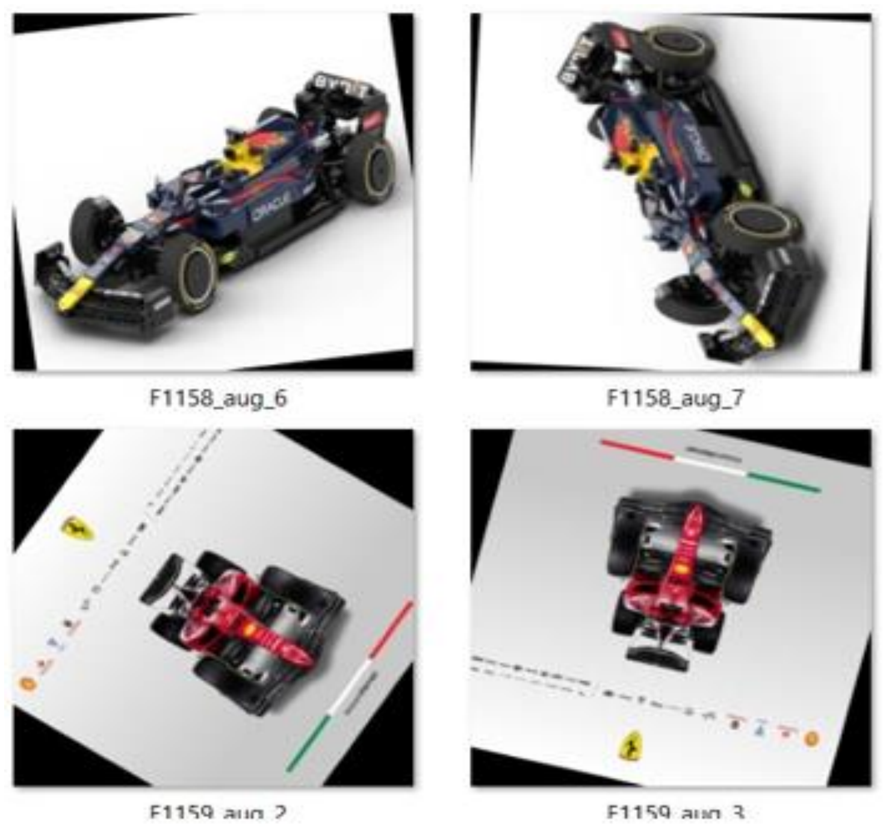
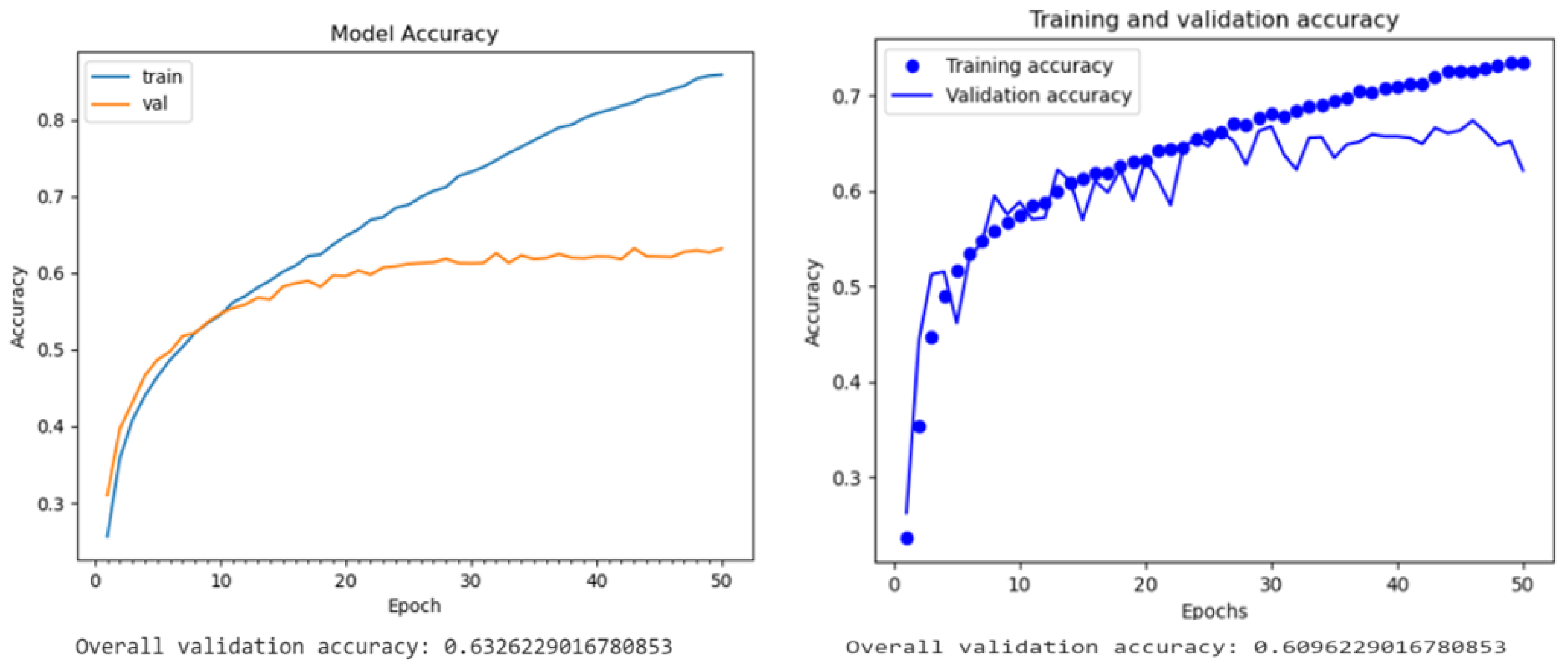


Figure 3 : Data Post Processing for YOLOv5

Results

The culmination of the project yields a comprehensive and diverse dataset, tailored specifically for various deep learning and AI endeavors, emphasizing object-centricity and richness in data representation.

Our generated datasets demonstrates superior performance in terms of accuracy percentages compared to existing online datasets. Through rigorous testing for emotion detection and YOLOv5 detection, our solution consistently outperforms competitors, yielding higher accuracy rates. This improvement translates to a better accuracy in AI models , and projects.



```
Overall validation accuracy: 0.6326229016780853
```

```
Overall validation accuracy: 0.6096229016780853
```

Figure 4 : Comparing model accuracy between our proprietary dataset and publicly available datasets.

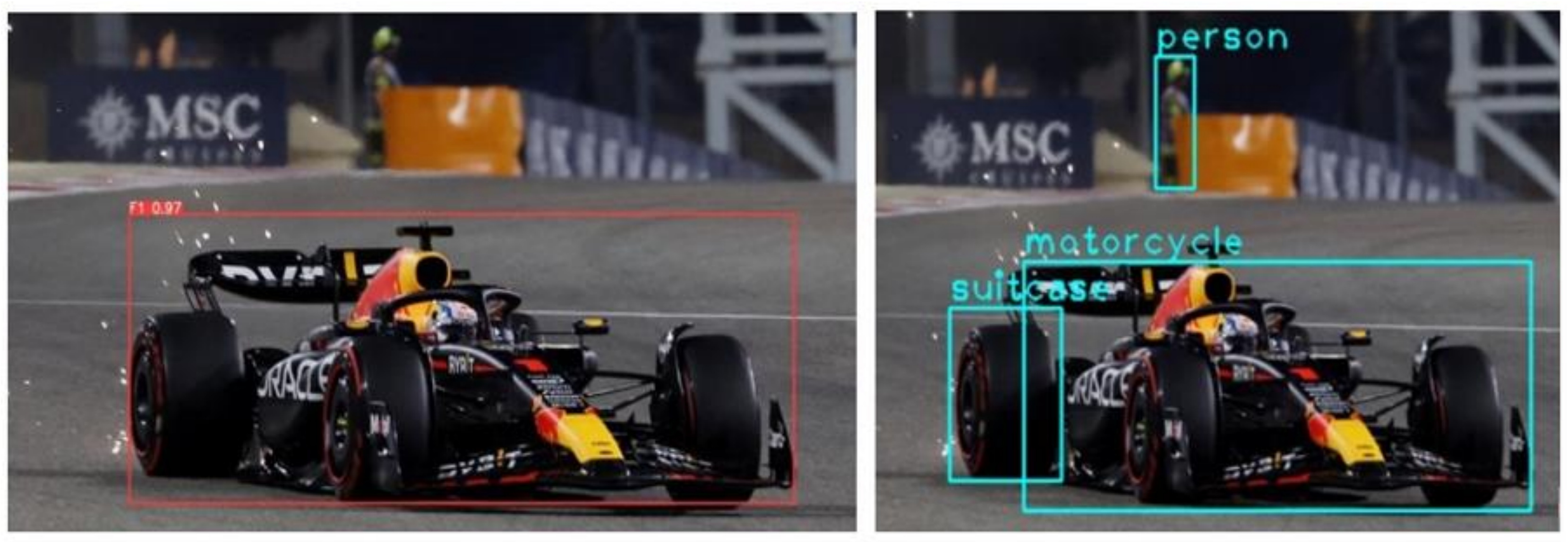
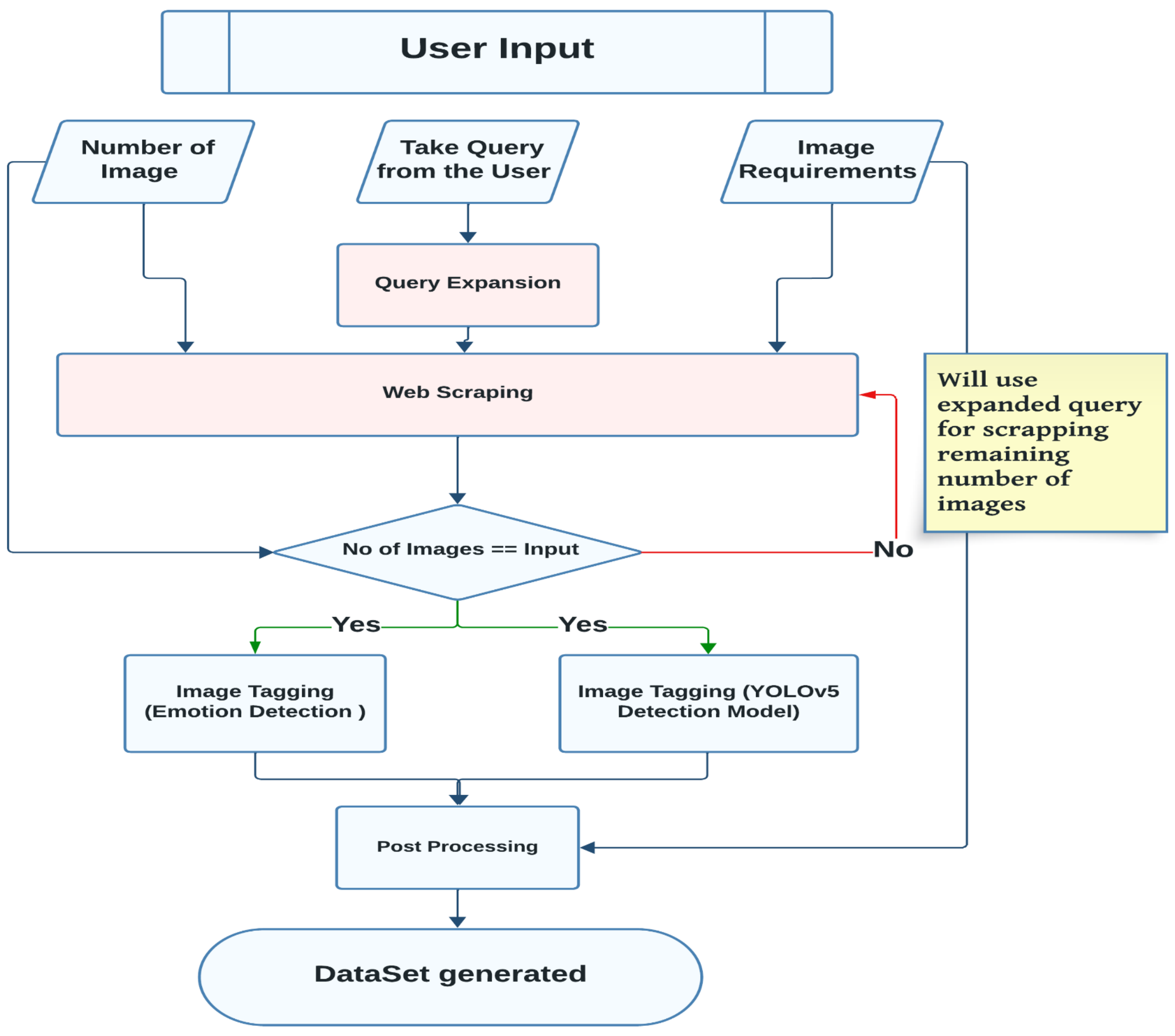


Figure 5 : Comparing model accuracy between our proprietary & publicly available datasets.



Conclusion

Our project represents a significant leap forward in AI, particularly in computer vision and emotion detection. Leveraging NLP for query expansion and automating dataset creation via web scraping, we've addressed the critical need for diverse, high-quality datasets. This streamlined approach ensures adherence to requirements, enhancing accuracy. Through image augmentation, we've bolstered dataset richness and model robustness. Future work involves updating Selenium web scraping scripts to adapt to changes in HTML structures and considering advancements in search engine behavior, potentially eliminating the need for noise-checking code. This work lays groundwork for innovation, promising real-world impact and pushing AI boundaries.

References

I. Dikmans, B., & Kang, D. (2023). A Brief Survey into the Field of Automatic Image Dataset Generation through Web Scraping and Query Expansion.

II. Niu, Qingli, et al. "Web Scraping Tool For Newspapers And Images Data Using Jsonify." Journal of Applied Science and Engineering 26.4 (2022): 465-474