Understanding Dataset Collection: (Problem-1)

ডেটাসেট তৈরির আগে একটা জিনিস একদম ক্লিয়ার করতে হবে—আমি ডেটা কালেক্ট করছি কেন? উদ্দেশ্যটা কী? এই প্রশ্নের উত্তর স্পষ্ট না হলে, পুরো ডেটা কালেকশন প্রসেসটাই কনফিউজড হবে।

আমি যখন Bangla Handwritten OCR নিয়ে কাজ করতে গেলাম, তখন দেখলাম আমার সামনে ৩টা মেইন সমস্যা দাঁড়িয়ে আছে—

Problem 1: ভোকাবুলারি / টোকেনাইজার নিয়ে বড় একটা কনফিউশন

বাংলা একটা কমপ্লেক্স স্ক্রিপ্ট। যদি ক্যারেক্টার-লেভেলে টোকেনাইজ করি, তাহলে অনেক ইনএফিশিয়েন্ট হবে। ধরেন, আমার কাছে "সোনার বাংলা" লেখা একটা ইমেজ আছে। যদি ক্যারেক্টার-ওয়াইজ OCR করি, তাহলে আউটপুট আসবে এরকম:

(“ো”, “স”, “ন”, “া”, “র”, “ব”, “া”, “ং”, “ল”, “া”)

টেকনিকালি ঠিক, কিন্তু ভাষাগত দিক থেকে ঠিক না। এই টোকেনাইজেশন মেথড দিয়ে মডেল কোনো মীনিংফুল আউটপুট দিতে পারবে না।

সমাধান**: Grapheme-level Tokenization**

এই সমস্যার জন্য আমি যে এপ্রোচটা বেশি প্র্যাক্টিক্যাল মনে করি, সেটা হলো গ্রাফিম লেভেল টোকেনাইজেশন। এখানে আমরা ভিজুয়ালি যেটা "একটি ইউনিট" হিসেবে দেখি, সেটা-ই টোকেন হিসেবে ধরবো।

উদাহরণস্বরূপ, "সোনার বাংলা" কে যদি গ্রাফিম-লেভেলে ব্রেক করি, তাহলে আমরা পাবো:

[“সো”, “না”, “র”, “বা”, “ং”, “লা”]

এটা অনেক বেশি ভাষা-ফ্রেন্ডলি এবং হিউম্যান রিডেবল। এই এপ্রোচটা আমি নিজে তৈরি করিনি—অনেক আগেই কিছু রিসার্চার এ নিয়ে কাজ করেছে। নিচে তাদের কিছু লিংক/পেপার দিয়ে দিব।

তাহলে আমাদের **Vocabulary** কেমন হবে**?**

আমাদের OCR মডেলের টোকেন ভোকাবুলারি বানাতে হলে নিচের ক্যাটাগরি গুলো ফোকাস করতে হবে:

1. স্বরবর্ণ (Vowels): ১১টা বেসিক স্বরবর্ণ
2. ব্যঞ্জনবর্ণ (Consonants): ৩৯টা বেসিক ব্যঞ্জনবর্ণ
3. সংখ্যা (Numerals): বাংলা ডিজিটস – ০ থেকে ৯
4. যুক্তাক্ষর (Compound Graphemes):
    - কার, ফলা, চন্দ্রবিন্দু, বিসর্গ সহ শত শত কমন কম্পাউন্ড ফর্ম
    - প্রায় ৬০ থেকে ৪০০ টার মতো most commonly used Juktakkhor

Summary:

ডেটা কালেক্ট করার আগে "ভোকাবুলারি ডিজাইন" করা একটা অনেক বড় স্টেপ। কারণ ভোকাবুলারিই ঠিক করে দিবে, মডেল কিভাবে ভাষাটা বুঝবে। Grapheme-level approach না নিলে, Bangla OCR সিস্টেম কোনোদিন প্রোডাকশন-রেডি হবে না।

🔍 **Reference Papers (**রিসার্চ বেইসড প্রুভ**):**

1. Bangla Character Recognition System — The Deep Learning way (1/n) - Medium, accessed July 10, 2025, https://medium.com/analytics-vidhya/bangla-character-recognition-system-the-deep-learning-way-1-n-8671a33a7860
2. Grandmasters Series: Learning from the Bengali Character Recognition Kaggle Challenge, accessed July 10, 2025, https://developer.nvidia.com/blog/grandmasters-series-learning-from-bengali-character-recognition-kaggle-challenge/
3. Handwritten Bangla Basic and Compound character recognition using MLP and SVM classifier - Bohrium, accessed July 10, 2025, https://bohrium.dp.tech/paper/arxiv/1002.4040
4. (PDF) A Large Multi-target Dataset of Common Bengali Handwritten ..., accessed July 10, 2025, https://www.researchgate.net/publication/354358208_A_Large_Multi-target_Dataset_of_Common_Bengali_Handwritten_Graphemes