এনএলপি পাইপলাইন হলো একটি পূর্ণাঙ্গ এনএলপি সফটওয়্যার তৈরির ধারাবাহিক ধাপসমূহের সমষ্টি। এটি কেবল কোড করা নয়, বরং একটি সমস্যাকে সমাধানের জন্য সঠিক চিন্তাধারা বা থিংকিং ফ্রেমওয়ার্ক তৈরি করতে সাহায্য করে। নিচে এনএলপি পাইপলাইনের পাঁচটি প্রধান ধাপ বিস্তারিতভাবে আলোচনা করা হলো:

১. ডেটা সংগ্রহ (Data Acquisition)

যেকোনো মেশিন লার্নিং বা এনএলপি প্রজেক্টের প্রথম এবং সবচেয়ে গুরুত্বপূর্ণ ধাপ হলো প্রয়োজনীয় ডেটা সংগ্রহ করা। ডেটা ছাড়া কোনো মডেল তৈরি করা অসম্ভব। ডেটা সংগ্রহের ক্ষেত্রে সাধারণত তিনটি পরিস্থিতি দেখা দেয়:

- নিজের কাছে ডেটা থাকলে: ডেটা যদি নিজের কম্পিউটারে বা কোম্পানির ডেটাবেসে থাকে, তবে কাজ সহজ। তবে ডেটা যদি খুব কম হয়, তখন ডেটা অগমেন্টেশন (Data Augmentation) ব্যবহার করে কৃত্রিমভাবে ডেটা বাড়ানো হয়। এর জন্য সিনোনিম (সমার্থক শব্দ) ব্যবহার করা, ব্যাক-ট্রান্সলেশন (এক ভাষা থেকে অন্য ভাষায় নিয়ে আবার আগের ভাষায় ফিরিয়ে আনা) অথবা গ্রামাটিক্যাল পরিবর্তন করে নতুন টেক্সট তৈরি করা হয়।
- ডেটা অন্যের কাছে থাকলে: অনেক সময় পাবলিক ডেটাসেট (যেমন- Kaggle) থেকে ডেটা নেওয়া হয়। আবার প্রয়োজনে কম্পিটিটরদের ওয়েবসাইট থেকে ওয়েব স্ক্র্যাপিং করে বা বিভিন্ন এপিআই (API) ব্যবহার করে ডেটা সংগ্রহ করা যায়। এমনকি পিডিএফ ফাইল, ছবি (OCR প্রযুক্তি ব্যবহার করে) বা অডিও ফাইল থেকেও টেক্সট বের করে আনা সম্ভব।
- ডেটা কারো কাছেই না থাকলে: এটি সবচেয়ে চ্যালেঞ্জিং পরিস্থিতি। এখানে শুরুতে খুব সামান্য ডেটা নিয়ে রুল-বেসড বা হিউরিস্টিক পদ্ধতিতে কাজ শুরু করতে হয় এবং সময়ের সাথে সাথে ব্যবহারকারীদের থেকে ফিডব্যাক নিয়ে ডেটাবেস বড় করতে হয়।

২. টেক্সট প্রিপারেশন (Text Preparation)

সংগৃহীত ডেটা সরাসরি মডেলে ব্যবহার করা যায় না কারণ এতে অনেক ভুল বা অপ্রয়োজনীয় তথ্য থাকে। টেক্সট প্রিপারেশনকে তিন ভাগে ভাগ করা যায়:

- বেসিক ক্লিনিং: টেক্সট থেকে অপ্রয়োজনীয় HTML ট্যাগ সরানো, ইমোজিগুলোকে মেশিনের বোঝার উপযোগী টেক্সটে রূপান্তর করা এবং বানানের ভুল সংশোধন করা।
- বেসিক প্রি-প্রসেসিং: এখানে টোকেনাইজেশন করা হয়, অর্থাৎ পুরো টেক্সটকে ছোট ছোট শব্দ বা বাক্যে ভাগ করা হয়। এরপর 'the', 'is', 'and'-এর মতো স্টপ-ওয়ার্ডস বাদ দেওয়া হয় কারণ এগুলোর বিশেষ কোনো অর্থ থাকে না। এছাড়াও শব্দের মূল রূপ বের করার জন্য স্টেমিনিং বা লেমাটাইজেশন করা হয় এবং সব টেক্সটকে লোয়ার-কেস (ছোট হাতের অক্ষর) করা হয়।
- অ্যাডভান্সড প্রি-প্রসেসিং: চ্যাটবট বা জটিল অ্যাপ্লিকেশনের জন্য পার্টস অফ স্পিচ (POS) ট্যাগিং এবং কোরফারেন্স রেজোলিউশন (যেমন: 'সে' বলতে আগের বাক্যের কোন ব্যক্তিকে বোঝানো হচ্ছে তা চিহ্নিত করা) করা হয়।

কোরফারেন্স রেজোলিউশন (Coreference Resolution) হলো ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং (NLP)-এর একটি প্রক্রিয়া, যার মাধ্যমে একটি টেক্সটে ভিন্ন ভিন্ন শব্দ বা শব্দগুচ্ছ (যেমন সর্বনাম, নাম, বা বিশেষ্য পদ) একই ব্যক্তি বা বস্তুকে নির্দেশ করছে কিনা তা শনাক্ত করা হয়। এর মূল লক্ষ্য হলো, কোনো একটি প্রসঙ্গে ব্যবহৃত বিভিন্ন উল্লেখ (mentions) যেমন 'তিনি', 'তার', 'এই ব্যক্তি', 'জনাব করিম'—এগুলো যে একই সত্তাকে বোঝাচ্ছে, তা চিহ্নিত করে

একটি 'কোরফারেন্স চেইন' বা গ্রুপ তৈরি করা, যা কম্পিউটারের পক্ষে মানুষের ভাষা বোঝা এবং তথ্য বিশ্লেষণ (যেমন সারসংক্ষেপ তৈরি, প্রশ্নোত্তর) করার জন্য অত্যন্ত জরুরি।

NLP-তে Parts of Speech (POS) Tagging হলো প্রতিটি শব্দকে তার ব্যাকরণগত শ্রেণি (যেমন বিশেষ্য, ক্রিয়া, বিশেষণ) অনুযায়ী চিহ্নিত করার প্রক্রিয়া, যা বাক্য থেকে অর্থ ও সম্পর্ক বুঝতে সাহায্য করে এবং মেশিন ট্রান্সলেশন, সেন্টিমেন্ট অ্যানালাইসিস, স্প্যাম ডিটেকশনের মতো কাজে ব্যবহৃত হয়

৩. ফিচার ইঞ্জিনিয়ারিং (Feature Engineering)

কম্পিউটার বা মেশিন লার্নিং অ্যালগরিদম সরাসরি টেক্সট বুঝতে পারে না, তারা শুধু সংখ্যা বোঝে। তাই টেক্সটকে সংখ্যায় রূপান্তর করার প্রক্রিয়াই হলো ফিচার ইঞ্জিনিয়ারিং। একে টেক্সট ভেক্টরাইজেশনও বলা হয়।

- সাধারণ পদ্ধতি: Bag of Words, TF-IDF বা Word2Vec-এর মতো টেকনিক ব্যবহার করে শব্দকে ভেক্টরে রূপান্তর করা হয়।
- মেশিন লার্নিং বনাম ডিপ লার্নিং: মেশিন লার্নিং মডেলে ইঞ্জিনিয়ারকে নিজের জ্ঞান বা ডোমেইন নলেজ ব্যবহার করে ফিচার তৈরি করতে হয়। অন্যদিকে, ডিপ লার্নিং মডেলে মডেল নিজেই অটোমেটিকভাবে ফিচার তৈরি করে নেয়, যদিও এটি কীভাবে কাজ করে তা বোঝা অনেক সময় কঠিন হয়ে পড়ে।

৪. মডেলিং ও মূল্যায়ন (Modeling & Evaluation)

ডেটা যখন সংখ্যায় রূপান্তরিত হয়, তখন সেটির ওপর অ্যালগরিদম প্রয়োগ করা হয়।

- মডেলিং: ডেটার পরিমাণ এবং সমস্যার ধরন অনুযায়ী হিউরিস্টিক (নিয়ম-ভিত্তিক), মেশিন লার্নিং বা ডিপ লার্নিং পদ্ধতি বেছে নেওয়া হয়। বর্তমানে ট্রান্সফার লার্নিং (আগে থেকে অন্য বড় ডেটায় প্রশিক্ষিত মডেল ব্যবহার করা) খুবই জনপ্রিয়।
- মূল্যায়ন: মডেলটি কেমন কাজ করছে তা দুইভাবে দেখা হয়। ইনট্রিনসিক (Intrinsic) মূল্যায়নে টেকনিক্যাল মেট্রিক্স যেমন Accuracy বা Precision দেখা হয়। আর এক্সট্রিনসিক (Extrinsic) মূল্যায়নে দেখা হয় বাস্তব ব্যবসায়িক ক্ষেত্রে এটি কতটা প্রভাব ফেলছে (যেমন: গুগল কিবোর্ডে সাজেস্ট করা শব্দ মানুষ আসলেই কতবার সিলেক্ট করছে)।

৫. ডিপ্লয়মেন্ট (Deployment)

সবশেষে মডেলটিকে ব্যবহারকারীদের জন্য উন্মুক্ত করা হয়।

- ডিপ্লয় ও মনিটরিং: মডেলটিকে একটি মাইক্রো-সার্ভিস বা এপিআই হিসেবে ক্লাউড সার্ভারে রাখা হয়। এরপর একটি ড্যাশবোর্ডের মাধ্যমে সবসময় মনিটর করা হয় যে মডেলটি ঠিকমতো রেজাল্ট দিচ্ছে কি না।
- আপডেট: সময়ের সাথে সাথে মানুষের কথা বলার ধরন বা ডেটা বদলে গেলে মডেলটিকে পুনরায় নতুন ডেটা দিয়ে ট্রেন করিয়ে আপডেট করতে হয়।

একটি উদাহরণ: আপনি যদি কোরা (Quora)-র মতো একটি সিস্টেমে দুটি প্রশ্ন একই কি না তা ধরতে চান, তবে আপনাকে প্রথমে ডেটা সংগ্রহ করতে হবে, এরপর প্রশ্নগুলো ক্লিন করতে হবে, সেগুলোকে সংখ্যায় রূপান্তর করতে হবে এবং সবশেষে একটি মডেল তৈরি করে দেখতে হবে সেটি কতটা নির্ভুলভাবে ডুপ্লিকেট প্রশ্ন ধরতে পারছে।

পুরো এই প্রসেসটি সবসময় সরলরেখায় চলে না; অনেক সময় ডিপ্লয়মেন্টের পর রেজাল্ট খারাপ আসলে আবার ডেটা সংগ্রহ বা ফিচার ইঞ্জিনিয়ারিং ধাপে ফিরে যেতে হয়। একেই বলা হয় নন-লিনিয়ার প্রসেস।