ন্যাচারাল ল্যাঙ্গুয়েজ প্রোসেসিং (NLP)

লেখা: জিবরান জারিফ অমিয় · ১৪ অক্টোবর, ২০২২

NLP কি?

ন্যাচারাল ল্যাঙ্গুয়েজ প্রোসেসিং (Natural Language Processing) বা এনএলপি (NLP) হলো কৃত্রিম বুদ্ধিমত্তার (Artificial Intelligence) এমন একটি শাখা যা কম্পিউটারকে মানুষের লিখিত এবং কথ্য উভয় ভাষা বুঝতে এবং বিশ্লেষণে সক্ষম করে।

কেন NLP?

সৃজনশীল কল্পনা থেকে গল্প বলার ক্ষমতা, চিন্তাভাবনা, আকাঙ্ক্ষা এবং স্বপ্নকে কথায় এবং লেখায় প্রকাশ করার ক্ষমতা মানুষকে আলাদা এবং বিশেষ করে তোলে। যোগাযোগের সহজাত এই বৈশিষ্ট্যটি আমাদের একতাবদ্ধ হতে সাহায্য করেছে এবং কালের পরিক্রমায় কোটি কোটি ব্যক্তিকে একত্রে কাজ করতে এবং কাজের সমন্বয় রক্ষা করতে সক্ষম করেছে। ভাষা যোগাযোগের সবচেয়ে গুরুত্বপূর্ণ মাধ্যম। বর্তমান পৃথিবীতে সাত হাজারেরও বেশি ভাষা বিরাজমান। স্বাভাবিকভাবেই আমরা চাই কেবল মানুষই নয় বরং মানুষের উদ্ভাবিত প্রযুক্তিসমূহ যেমন কম্পিউটারও যেন মানুষের প্রাকৃতিক ভাষা বুঝতে পারে। সাধারণ কম্পিউটার মানুষের প্রাকৃতিক ভাষা বুঝতে সক্ষম নয়। এতদিনও কম্পিউটারের সাথে যোগাযোগ করার জন্য একজন ব্যবহারকারী বিভিন্ন প্রোগ্রামিং ল্যাঙ্গুয়েজ (Programming Language) যেমন জাভা (Java), সি/সি-প্লাস-প্লাস (C/C++), পাইথন (Python) ইত্যাদির সহায়তা নিতো। কিন্তু NLP এমন সব সফটওয়্যার তৈরি করবে যা কম্পিউটার অ্যালগরিদম (Algorithm) এবং কৃত্রিম বুদ্ধিমত্তা ব্যবহার করে মানুষের প্রাকৃতিক ভাষা চিহ্নিতকরণ, বিশ্লেষণ এবং প্রক্রিয়াকরণ করতে সক্ষম হবে। এতে মানুষ অনায়াসেই একটি কম্পিউটারের সাথে সংযোগ স্থাপন করতে পারবে কোনো ধরণের প্রোগ্রামিং ছাড়া। NLP একটি কম্পিউটারকে যথেষ্ট বুদ্ধিমত্তা এবং ভাষাগত দক্ষতা প্রদান করবে যাতে মনে হয় যেন আপনি একটি কম্পিউটার না, বরং একজন মানুষের সাথেই কথোপকথন বা ইন্টার‍্যাক্ট করছেন। NLP-র খুব পরিচিত দুটো অ্যাপ্লিকেশন হলো গুগল অ্যাসিস্ট্যান্ট (Google Assistant) এবং সিরি (Siri)।

NLP-র ইতিহাস

মানুষের প্রাকৃতিক ভাষা তৈরি, বিশ্লেষণ এবং বুঝতে সক্ষম এমন সফটওয়্যার উদ্ভাবনের প্রয়াসেই মূলত আবির্ভাব ঘটে NLP-র। ন্যাচারাল ল্যাঙ্গুয়েজ প্রোসেসিং-এর অধ্যয়ন শুরু হয় ১৯৫০-এর দশকে, যদিও কিছু কাজ আগের সময় থেকে পাওয়া যায়। ১৯৫০ সালে, অ্যালান টুরিং (Alan Turing) "কম্পিউটিং মেশিনারি অ্যান্ড ইন্টেলিজেন্স" (Computing Machinery and Intelligence) শিরোনামে একটি নিবন্ধ প্রকাশ করেছিলেন। মানুষের সমান বা মানুষ থেকে পৃথকযোগ্য নয় এমন বুদ্ধিমান আচরণ প্রদর্শন করার ক্ষমতাসম্পন্ন মেশিনের একটি পরীক্ষার কথা উল্লেখ করেন তিনি। এই পরীক্ষাটিই বর্তমানে **টিউরিং টেস্ট** (Turing Test) হিসেবে বিশ্বব্যাপী সুপরিচিত। সেই সময় প্রাকৃতিক ভাষা প্রক্রিয়াকরণকে **মেশিন ট্রান্সলেশন** (Machine Translation) বলা হতো। মেশিন ট্রান্সলেশন বিধিবদ্ধ কিছু নিয়ম দ্বারা ভাষাকে বেশ সরলভাবে প্রক্রিয়া করতো। ভাষা অনুবাদের লক্ষ্যে এটি অভিধান থেকে উপযুক্ত শব্দ খুঁজে আনতো এবং বাক্যের ব্যাকরণগত বৈধতা ঠিক রাখতে শব্দগুলোকে পুনর্বিন্যাস করে নিতো। সহজ সরল বাক্যানুবাদ যা নিয়মের বহির্ভূত নয় তা ঠিক থাকলেও, বাস্তবে মেশিন ট্রান্সলেশনের অনুবাদে প্রচুর ত্রুটি ও ভাষাগত অস্পষ্টতা প্রকট হয়ে ওঠে। ১৯৫৭ সালে চমস্কি (Chomsky) জেনারেটিভ গ্রামারের (Generative Grammar) ধারণা প্রবর্তন করেন যা মেশিন ট্রান্সলেশনের সাথে ভাষাবিজ্ঞানের (Linguistics) সমন্বয় সাধন করে। এতে মেশিন ট্রান্সলেশনের সঠিকতা বৃদ্ধি পায়। কিন্তু এক দশকেরও বেশি গবেষণা এবং বহু মিলিয়ন ডলার ব্যয়ের পরে, ১৯৬৬ সাল নাগাদ বোঝা গেল মেশিন ট্রান্সলেশন দ্বারা অনুবাদ সাধারণ মানব অনুবাদকের অনুবাদের চেয়েও বেশি ব্যয়বহুল ছিল কেননা কম্পিউটারের গণনাশক্তির বিকাশ তখনও কেবল প্রাথমিক স্তরে। তাই নূন্যতম কথোপকথন চালিয়ে যেতে সক্ষম এমন কম্পিউটারের কথা চিন্তা করাও ছিল দুষ্কর। এরপর মেশিন ট্রান্সলেশন এবং NLP-র গবেষণায় কিছুটা ভাটা পড়ে। ৭০এর দশকে গবেষকগণ বিভিন্ন ভাষাতত্ত্বের ওপর ব্যাপক গবেষণা চালান এবং ব্যাকরণগত শব্দক্রম ঠিক রেখে বাক্য থেকে প্রাসঙ্গিক অর্থ উদ্ধারের এমন সব পন্থা আবিষ্কার করেন যা কম্পিউটারের জন্য গণনামূলকভাবে এবং সময়ের দিক দিয়ে পূর্বের তুলনায় সাশ্রয়ী। এরই সূত্র ধরে আবির্ভাব ঘটে ELIZA, SHRDLU, LUNAR, PARRY-র মতো অসংখ্য NLP প্রোজেক্ট। ELIZA একজন মনোবিজ্ঞানী এবং একজন রোগীর কথোপকথন অনুকরণ করতে পারতো। PARRY ব্যক্তির ধারণা, মতামত, বিশ্বাস ইত্যাদি ইনপুটের ভিত্তিতে প্যারানয়েড সিজোফ্রেনিয়ায় আক্রান্ত একজন রোগীর অনুকরণ করতে পারতো। তদুপরি, ১৯৮০-এর দশক পর্যন্ত, বেশিরভাগ NLP সিস্টেমগুলি ছিল গৎবাঁধা বিধিবদ্ধ কিছু জটিল নিয়মের সমষ্টি। ১৯৮০-এর দশকের শেষের দিকে, কম্পিউটারের গণনা শক্তির বৃদ্ধিপ্রাপ্তি এবং ভাষা প্রক্রিয়াকরণের জন্য মেশিন লার্নিং (Machine Learning) অ্যালগরিদম প্রবর্তনের মাধ্যমে NLP-র একটি বিপ্লব ঘটেছিল। ১৯৯০-এর দশকের মাঝামাঝি সময় পর্যন্তও তথাকথিত পরিসংখ্যান বিপ্লবের (Statistical Revolution) পরেও NLP অনেকাংশেই মেশিন লার্নিং-এর উপর নির্ভরশীল ছিল। ২০০০-এর দশকে, ভাষা প্রক্রিয়াকরণের জন্য নিউরাল নেটওয়ার্ক (Neural Network) ব্যবহৃত হয়, যার লক্ষ্য ছিল পূর্ববর্তী শব্দগুলি বিশ্লেষণের মাধ্যমে পরবর্তী শব্দ প্রেডিক্ট করা। ২০১০-এর দিকে, **ডিপ লার্নিং**-এর (Deep Learning) আবির্ভাব ঘটে। ওয়ার্ড এম্বেডিং (Word Embedding)-এর মাধ্যমে অর্থ, প্রাসঙ্গিকতা বা

প্রভাবের ভিত্তিতে অজস্র শব্দের সম্পর্ক নির্ণয় করা সম্ভব হয়। নিউরাল নেটওয়ার্কের বিভিন্ন শাখার বিকাশ ঘটে যেমন রিকারেন্ট নিউরাল নেটওয়ার্ক (RNN), কনভোলিউশনাল নিউরাল নেটওয়ার্ক (CNN), রিকার্সিভ নিউরাল নেটওয়ার্ক ইত্যাদি। RNN ব্যবহারে NLP মডেলগুলোকে পরিবর্তনশীল ইনপুটের জন্য যথাযথ আউটপুট দেওয়ার জন্য প্রস্তুত করা হয়। ২০১৪ সালে **সিকুয়েন্স-টু-সিকুয়েন্স** (Sequence-to-Sequence) লার্নিং-এর আগমনের পর NLP নিউরাল নেটওয়ার্ক ব্যবহার করে ভাষান্তর করা কিংবা একটি প্যারাগ্রাফ, ছবি বা ভয়েস ইনপুট থেকে সারমর্ম, মূল বক্তব্য বা প্রাসঙ্গিক অর্থ উদ্ধারে সক্ষম হয়। ডিপ নিউরাল নেটওয়ার্ক (Deep Neural Network) সংযোজিত মেশিন লার্নিং পদ্ধতিগুলো ভাষা প্রক্রিয়াকরণে ব্যাপকতা লাভ করে এবং দ্রুত এবং সময় সাশ্রয়ী উপায়ে অত্যাধুনিক ফলাফল অর্জনে সক্ষম হয়। বর্তমানে, ডিপ লার্নিং-এর সহায়তায় সম্ভব হচ্ছে টেক্সট (Text) বা লেখা (শব্দের ক্রম বা অক্ষরের ক্রম যা মানব ভাষার ভিত্তি তৈরি করে) প্রক্রিয়াকরণের মাধ্যমে ডকুমেন্ট ক্লাসিফিকেশন, টাইম সিরিজ বিভাজন (Time Series Classification) যেমন আবহাওয়ার পূর্বাভাস, স্টক মার্কেটের মূল্য প্রেডিকশন, সিকুয়েন্স-টু-সিকুয়েন্স লার্নিং ব্যবহার করে একটি ইংরেজি বাক্যকে বাংলায় ভাষান্তর করা ইত্যাদি।

NLP কিভাবে কাজ করে?

NLP মূলত মেশিন লার্নিং মডেলের মাধ্যমে মানুষের প্রাকৃতিক ভাষা প্রক্রিয়াকরণ করে থাকে। এই মডেলটি টেক্সট বা লেখা আকারে ডেটা গ্রহণ করে এবং ধারাবাহিক কিছু প্রক্রিয়ার মাধ্যমে সেই টেক্সট বা লেখার অর্থ উদ্ঘাটন করে থাকে। NLP-এর আবার কিছু উপসেট রয়েছে। একটি হলো **NLU** বা ন্যাচারাল ল্যাঙ্গুয়েজ আন্ডারস্ট্যান্ডিং যা প্রদত্ত লেখা বা ইনপুট টেক্সট থেকে অর্থ উদ্ধার করে এবং বিভিন্ন বৈশিষ্ট্যের ভিত্তিতে টেক্সটের বিভিন্ন অংশকে শ্রেণিবদ্ধ বা কাঠামোবদ্ধ করে যা কম্পিউটারের জন্য বোধগম্য হয়। আরেকটি হলো **NLG** বা ন্যাচারাল ল্যাঙ্গুয়েজ জেনারেশন যা কাঠামোবদ্ধ ডেটা থেকে প্রাকৃতিক ভাষা তৈরি করে যা মানুষের বোধগম্য হয়। NLP, NLU বা NLG বিভিন্ন টেকনিক ব্যবহার করে প্রাকৃতিক ভাষা প্রক্রিয়াকরণ করে থাকে। নিম্নে গুরুত্বপূর্ণ কয়েকটি টেকনিক আলোচনা করা হলো:

সেগমেন্টেশন (Segmentation):

সেগমেন্টেশন হলো কোনো ডকুমেন্টের যাবতীয় লেখা, রচনা বা প্যারাগ্রাফ (Paragraph) থেকে সেন্টেন্স (Sentence) বা বাক্যগুলোকে আলাদা করা। বিষয়টি আপাতদৃষ্টিতে বেশ সহজ মনে হতে পারে। নির্দিষ্ট কোনো বিরাম চিহ্ন বা পাঙ্কচুয়েশন মার্ক (Punctuation Mark) দ্বারা বাক্যগুলোকে পৃথক করলেই হলো। যেমন, বাংলা ভাষায় বাক্য আলাদা করা যায় দাঁড়ি (।) দেখে, আবার ইংরেজিতে সেন্টেন্স আলাদা করা হয় ফুল-স্টপ (.) দ্বারা। কিন্তু এর কিছু জটিলতাও আছে। অনেকসময় কেবল একটি দাঁড়ি বা ফুল-স্টপ দিয়ে বাক্য আলাদা করা যায় না।

মধ্যযুগীয় বাংলা সাহিত্যে আমরা দুই দাঁড়ির (।।) উপস্থিতি লক্ষ করি। আধুনিক যুগে এসেও অনেক কবি সাহিত্যিক কবিতার সমাপ্তি বোঝাতে দুই দাঁড়ি ব্যবহার করেন। ইংরেজিতে আবার ফুল-স্টপের বহুমাত্রিক ব্যবহার দেখা যায়। শব্দকে সংক্ষিপ্ত আকারে লেখার জন্য ফুল-স্টপ ব্যবহার করা হয়। যেমন, Mr., Mrs., Etc. ইত্যাদি। অর্থাৎ একটি প্যারাগ্রাফ থেকে বাক্য পৃথকীকরণের জন্য শুধু একটি দাঁড়ি বা ফুল-স্টপ ছাড়াও বেশ কিছু বিষয়ের প্রতি লক্ষ রাখা প্রয়োজন। সেগমেন্টেশনের উদাহরণ দেখে নেওয়া যাক।

“Sonargaon has a museum bearing witness of the old regime of Bengal. Lok Shilpa Jadughar of Sonargaon was established by the renowned painter Joynul Abedin in 1975. It will cost you 10 takas to enter the museum area. It’s a vast area, and will take you lots of time to round the entire place.”

সেগমেন্টেশনের পর প্যারাগ্রাফটি চারটি পৃথক বাক্যে পরিণত হবে।

1. Sonargaon has a museum bearing witness of the old regime of Bengal.
2. Lok Shilpa Jadughar of Sonargaon was established by the renowned painter Joynul Abedin in 1975.
3. It will cost you 10 takas to enter the museum area.
4. It’s a vast area, and will take you lots of time to round the entire place.

টেক্সট প্যারাগ্রাফ থেকে বাক্যে রূপান্তর হয়ে গেল। এখন প্রতিটি বাক্য থেকে আবার শব্দ আলাদা করতে হবে।

টোকেনাইজেশন (Tokenization):

একটি বাক্য থেকে ওয়ার্ড (Word) বা শব্দ পৃথক করার প্রক্রিয়াই হলো টোকেনাইজেশন। প্রক্রিয়াটি অনেকটাই সেগমেন্টেশনের মতো, যেখানে একটি চিহ্নের ভিত্তিতে উপাদানগুলো আলাদা করা হয়। সেগমেন্টেশনে যেমন দাঁড়ি বা ফুল-স্টপের ভিত্তিতে প্যারাগ্রাফ থেকে বাক্য আলাদা করা হয়েছিল, তেমনিভাবে টোকেনাইজেশনে বাক্য থেকে শব্দ পৃথক করা হয় স্পেস (Space) চিহ্নিত করে। তবে এখানেও ব্যতিক্রম আছে। কিছু শব্দ কেবল একটি স্পেস দ্বারা আলাদা করা যায় না। ইংরেজি কম্পাউন্ড নাউন (Compound Noun) বা যৌগিক বিশেষ্য যেমন, Mother-in-law, Blackboard, Football ইত্যাদি শব্দ দুই বা ততোধিক শব্দের সমাহার। এসকল শব্দকে আলাদাভাবে বিবেচনা করতে হয়। উপরের সেগমেন্টেশনের উদাহরণ থেকে প্রথম বাক্যটিকে টোকেনাইজ (Tokenize) করলে এই শব্দগুলো পাওয়া যায়, ‘Sonargaon’, ‘has’, ‘a’, ‘museum’, ‘bearing’, ‘witness’, ‘of’, ‘the’, ‘old’, ‘regime’, ‘of’, ‘Bengal’। উল্লেখ্য, এখানে প্রতিটি শব্দকে বলা হয় একেকটি **ইউনিগ্রাম** (Unigram)। টোকেনাইজেশনের মাধ্যমে আমরা নতুন যে উপাদানগুলো পেয়ে থাকি তাদের টোকেন বলে। টোকেন সাধারণত

একটি শব্দ হয়ে থাকে। যখন টোকেনে কেবল একটি শব্দ উপস্থিত থাকে তাকে ইউনিগ্রাম বলে। কখনও একটি টোকেনে একাধিক শব্দ থাকতে পারে। যেমন, 'Joynul Abedin' একটি **বাইগ্রাম** (Bigram) কেননা এতে দুটো শব্দ রয়েছে। আবার, তিনটি শব্দ থাকায় 'It will cost' টোকেনটি একটি **ট্রাইগ্রাম** (Trigram)। একইভাবে, টোকেনে যতগুলো শব্দ থাকে তাকে ততো গ্রাম (N-gram) বলা হয়।

স্টেমিং (Stemming):

স্টেম (Stem) শব্দের অর্থ হলো 'মূল'। একটি শব্দ থেকে উপসর্গ (Prefix) এবং প্রত্যয় (Suffix) অপসারণের মাধ্যমে ঐ শব্দের শব্দমূল নির্ধারণ করার প্রক্রিয়াকেই বলা হয় স্টেমিং। যেমন, 'like', 'likes', 'likely' প্রত্যেকটি শব্দেরই শব্দমূল কিন্তু একই। শব্দমূলটি হলো 'like'। অর্থাৎ একটি NLP মডেল স্টেমিং-এর মাধ্যমে এই তিনটি শব্দকে একই বিবেচনা করবে। শব্দের বিভিন্ন রূপ থাকা সত্ত্বেও স্টেমিং শব্দগুলোকে একীভূত করতে এবং অপ্রয়োজনীয় শব্দসমূহ অপসারিত করতে সক্ষম। আধুনিক সার্চ ইঞ্জিনগুলো স্টেমিং ব্যবহার করে আরও উন্নত এবং প্রাসঙ্গিক ফলাফল প্রদান করছে। স্টেমিং আবিষ্কারের পূর্বে 'fish' সার্চ করলে আক্ষরিক অর্থে 'fish' আছে এমন ওয়েবসাইটগুলো রেকমেন্ড করা হতো। কিন্তু 'fishes' বা 'fishing' শব্দ দুটো প্রাসঙ্গিক হওয়া সত্ত্বেও বাদ পরে যেত। স্টেমিং এই সমস্যার সমাধান করেছে এবং শব্দগুলোকে আলাদাভাবে বিবেচনা না করে বরং এদের শব্দমূল দ্বারা সম্পর্কিত করে করেছে। পর্টারস স্টেমার অ্যালগরিদম (Porter's Stemmar Algorithm) সবচেয়ে জনপ্রিয় স্টেমিং পদ্ধতিগুলোর একটি যা ১৯৮০ সালে প্রস্তাবিত হয়েছিল। এটি এই ধারণার উপর প্রতিষ্ঠিত যে ইংরেজি ভাষায় শব্দসমূহ বিভিন্ন প্রত্যয় এবং উপসর্গের সংমিশ্রণে তৈরি যা সহজেই বিভাজন করা যায়। শব্দ থেকে কেবল প্রত্যয় এবং উপসর্গগুলো সরাসরি অপসারণ করলেই শব্দমূল পাওয়া যায়। কিন্তু এই ধারণার কিছু সীমাবদ্ধতা রয়েছে। যেমন, 'ability' শব্দটি স্টেমিং করার পর দাঁড়াবে 'abil'। অথচ এর সঠিক শব্দমূল হলো 'able'। স্টেমিং সংক্রান্ত খুবই পরিচিত দুটি সমস্যা হলো **ওভার-স্টেমিং** (Over-Stemming) এবং **আন্ডার-স্টেমিং** (Under-Stemming)। ওভার-স্টেমিং হলো যখন সম্পর্ক নেই এমন শব্দগুচ্ছের একই শব্দমূল নির্ধারণ করা। যেমন, 'universal', 'university', 'universe' শব্দগুচ্ছের মূল ধরা হবে 'univers'। কিন্তু লক্ষ করুন শব্দ তিনটির কোনোই সম্পর্ক নেই, প্রতিটি শব্দই সম্পূর্ণ ভিন্ন কিছু বোঝায়। অর্থাৎ 'universe' লিখে সার্চ করলে মহাবিশ্বের পাশাপাশি 'university' বিষয়ক ওয়েবসাইটও রেকমেন্ড করা হবে যা পুরোপুরি অপ্রাসঙ্গিক। আবার ধরুণ, 'better' এবং 'good'। শব্দদ্বয় যদিও সম্পর্কিত স্টেমিং এদের আলাদা বিবেচনা করবে কেননা উভয়ের মূল ভিন্ন। এই সমস্যাকে বলা হয় আন্ডার-স্টেমিং। তাহলে বোঝা গেল স্টেমিং যেহেতু শব্দগুচ্ছের আক্ষরিক মূল নির্ধারণ করে এটি প্রসঙ্গ বা কন্টেক্সট (Context) বুঝতে সক্ষম নয়। অর্থাৎ আক্ষরিক মূল দ্বারা সম্পর্কিত এমন শব্দগুচ্ছের ক্ষেত্রে স্টেমিং ব্যবহার করা গেলেও বাস্তবে এর প্রয়োগে ত্রুটিপূর্ণ ফলাফল পাওয়ার সম্ভাবনা আছে।

লেমাটাইজেশন (Lemmatization):

স্টেমিং-এর সীমাবদ্ধতা কাটিয়ে উঠতে এবং শব্দগুলোকে প্রাসঙ্গিকতার ভিত্তিতে শ্রেণিবদ্ধ বা সম্পর্কিত করার প্রক্রিয়াই হলো লেমাটাইজেশন। লেমাটাইজেশনে শব্দমূলগুলোকে লেমা (Lemma) বলা হয়। এটি সমৃদ্ধ শব্দভান্ডার (Vocabulary) ব্যবহার করে এবং **মরফোলজিক্যাল অ্যানালাইসিস** (Morphological Analysis)-এর মাধ্যমে প্রসঙ্গ ঠিক রেখে সম্পর্কিত শব্দসমূহকে একই একই শ্রেণিভুক্ত করে। মরফোলজিক্যাল অ্যানালাইসিস হলো ভাষাবিজ্ঞানের এমন একটি শাখা যা শব্দের গঠন অধ্যয়ন করে। এটি নির্ধারণ করে কিভাবে একটি মরফিম (Morpheme) থেকে শব্দ উৎপন্ন হয়। মরফিম ইংরেজি ভাষার একটি মৌলিক এবং ক্ষুদ্রতম একক যার একটি নির্দিষ্ট ব্যাকরণগত অর্থ রয়েছে। মরফিম প্রক্রিয়াকরণের মাধ্যমে শব্দের প্রাসঙ্গিক অর্থ, ক্রিয়ার কাল, ভাব ইত্যাদি স্পষ্ট হয়ে ওঠে। এভাবেই লেমাটাইজেশন শব্দগুচ্ছের সঠিক সম্পর্ক নির্ণয় করতে সক্ষম হয়। যেমন, 'universal', 'university', 'universe' শব্দগুলোকে লেমাটাইজেশন আলাদা বিবেচনা করে, যেখানে স্টেমিং একীভূত করে ফেলত। অপরপক্ষে, 'good', 'better', 'best' শব্দগুচ্ছ একই লেমা দ্বারা চিহ্নিত হয়। শব্দ তিনটির লেমা হবে 'good', কিন্তু স্টেমিং হলে এদের পৃথক করে দিত। আধুনিক সার্চ ইঞ্জিনগুলো লেমাটাইজেশন ব্যবহারের মাধ্যমেই প্রয়োজনীয় এবং প্রাসঙ্গিক বিষয়াদি উপস্থাপন করে।

পিওএস ট্যাগিং (POS Tagging):

প্রত্যেকটি শব্দেরই একটি ব্যাকরণগত অবস্থান রয়েছে। যেমন, 'Sonargaon' একটি বিশেষ্য (Noun), 'bearing' একটি ক্রিয়া (Verb), 'of' আবার অব্যয় (Preposition)। শব্দসমূহের ব্যাকরণগত ভিন্নতার প্রেক্ষিতে বিভিন্ন শ্রেণি বা গোত্রে বিভক্ত করা হয়। এই শ্রেণি বা গোত্রগুলোই একেকটি Parts-of-Speech। একটি শব্দ কোন Parts-of-Speech তা নির্ণয় করে, ঐ শব্দের সঙ্গে যুক্ত বা ট্যাগ (Tag) করে দেওয়া হয়। শব্দসমূহকে তাদের Part-of-Speech-এর সাথে ট্যাগ করে দেওয়ার এই প্রক্রিয়াই হলো POS (Parts-of-Speech) Tagging। POSগুলো সংক্ষিপ্ত আকারে চিহ্নিত করা হয়। যেমন, Noun হলে 'N', Verb হলে 'VB', Adjective হলে 'JJ' ইত্যাদি। অনেকসময়, POSগুলোকে আরও নির্দিষ্টকরণের লক্ষ্যে অতিরিক্ত কিছু চিহ্ন যুক্ত করা হয়। যেমন, 'NN' দিয়ে বোঝানো হয় Singular Noun, 'NNP' হলো Proper Noun, 'VBN' হলো Verbএর Past Participle রূপ, 'PRP' হলো Personal Pronoun ইত্যাদি। POS Tagging-এর প্রয়োগ দেখে নেওয়া যাক। 'Sonargaon', 'has', 'a', 'museum', 'bearing', 'witness', 'of', 'the', 'old', 'regime', 'of', 'Bengal' শব্দগুলোর উপর POS Tagging প্রয়োগ করলে পাওয়া যাবে, 'Sonargaon - NNP', 'has - VB', 'a - DT', 'museum - NN', 'bearing - VBG', 'witness - NN', 'of - IN', 'the - DT', 'old - JJ', 'regime - NN', 'of - IN', 'Bengal - NNP'। পূর্ববর্তী শব্দের ভিত্তিতে পরবর্তী শব্দ কি হতে পারে তা এই POS Tagging ব্যবহার করে আধুনিক NLP মডেলসমূহ প্রেডিক্ট (Predict) করে ফেলে। যেমন, কোনো Noun বা Pronoun এর পরে সাধারণত Verb আসে।

NLP মডেল প্রাসঙ্গিকতার প্রেক্ষিতে সম্ভাব্য Verbগুলো রেকমেন্ড করে। এভাবেই বিভিন্ন অটোকমপ্লিট (Autocomplete) সফটওয়্যার তৈরি করা হয়।

এনইআর (NER):

এতক্ষণ আমরা দেখেছি শব্দগুচ্ছের অর্থ বা প্রাসঙ্গিকতার ভিত্তিতে কিংবা একটি শব্দ কোন Parts-of-Speech তা নির্ণয়ের মাধ্যমে শব্দসমূহকে নির্দিষ্ট শ্রেণিভুক্ত বা সম্পর্কিত করা যায়। কিন্তু কিছু শব্দ ব্যাকরণগতভাবে একই Parts-of-Speech-এর আওতাধীন হওয়া সত্ত্বেও তাদের তাৎপর্য ভিন্ন। যেমন, 'Sonargaon', 'Joynul Abedin' এবং 'Bengal' শব্দ তিনটির POS কিন্তু একই। এদের প্রত্যেকটিই 'NNP' বা Proper Noun। কিন্তু খেয়াল করুন, 'Sonargaon' একটি স্থানের নাম, 'Joynul Abedin' একজন ব্যক্তির নাম, আবার 'Bengal' বলতে বোঝায় একটি ঐতিহাসিক ভূখণ্ড বা প্রদেশ। তাহলে কিছু শব্দকে একটি নির্দিষ্ট সত্ত্বা বা এনটিটি (Entity) দ্বারা চিহ্নিত করা যায়। এই প্রক্রিয়াটিই হলো NER (Named Entity Recognition)। NER প্রয়োগের মাধ্যমে NLP মডেলগুলো বুঝতে পারে একটি শব্দ কোন Entity নির্দেশ করে। NLP বিভিন্ন ধরণের Entity শনাক্তকরণে সক্ষম। যেমন, সংগঠন, পরিমাণ, আর্থিক মূল্যবোধ, শতাংশ বা Percentage, মানুষের নাম, কোম্পানির নাম, ভৌগলিক অবস্থান (ভৌতিক এবং রাজনৈতিক উভয়), পণ্যের নাম, তারিখ এবং সময়, টাকার পরিমাণ, ঘটনার নাম ইত্যাদি। 'It', 'will', 'cost', 'you', '10 takas', 'to', 'enter', 'the', 'museum', 'area' এর ওপর NER প্রয়োগের মাধ্যমে NLP বুঝে নেবে যে, 'you' বলতে একজন মানুষকে বোঝানো হচ্ছে, '10 takas' হলো টাকার একটি পরিমাণ এবং 'museum' হলো ভ্রমণের একটি স্থান। NER বাক্যে উপস্থিত শব্দগুচ্ছের প্রাসঙ্গিকতা নির্ণয়ে সহায়ক ভূমিকা পালন করে।

ব্যাগ অফ ওয়ার্ডস (Bag of Words):

ব্যাগ অফ ওয়ার্ডস বা বিওডব্লিউ (BoW) মডেলটি প্রাকৃতিক ভাষা প্রক্রিয়াকরণ এবং তথ্য পুনরুদ্ধার (Information Retrieval) কাজে ব্যবহৃত একটি সরলীকৃত পদ্ধতি। এটি এক বা একাধিক ডকুমেন্ট বা ফাইলে উপস্থিত অসংখ্য শব্দকে আলাদা করে। এবার প্রাপ্ত শব্দগুলোর ব্যাকরণ, শব্দক্রম ইত্যাদি বৈশিষ্ট্য উপেক্ষা করে সকল শব্দকে একটি থলে বা ব্যাগে ভরে নেয়। এরপর সেই ব্যাগে থাকা শব্দগুলোর পুনরাবৃত্তি বা রিপিটিশান (Repetition) বা ফ্রিকুয়েন্সি (Frequency) নির্ণয় করে। এর মাধ্যমে বোঝা যায় ডকুমেন্ট বা ফাইলগুলোয় কোন শব্দগুলোর উপস্থিতি বেশি। এতে একটি ডকুমেন্ট কি বিষয়ে আলোকপাত করছে তা স্পষ্ট হয়ে ওঠে। ব্যাগ অফ ওয়ার্ডসের একটি উদাহরণ দেখে নেওয়া যাক। ধরুন, একটি রেস্টুরেন্টের ম্যানেজার তাদের পরিবেশিত খাবার সম্পর্কে কাস্টমারদের মতামত যাচাই করতে চাচ্ছেন। এতে তিনি রেস্টুরেন্টের ওয়েবসাইটের রিভিউ (Review) সেকশন দেখা শুরু করলেন।

হাজার হাজার রিভিউ থেকে তিনি চারটি রিভিউ বাছাই করলেন। এখানে একেকটি রিভিউ একেকটি ডকুমেন্ট। রিভিউ চারটি নিম্নরূপ:

রিভিউ-১: This pasta is very tasty and affordable.

রিভিউ-২: This pasta is not tasty and is affordable.

রিভিউ-৩: This pasta is delicious and cheap.

রিভিউ-৪: Pasta is tasty and pasta tastes good.

এখন আমরা যদি চারটি রিভিউয়ের অনন্য বা ইউনিক (Unique) শব্দের সংখ্যা গণনা করি তাহলে আমরা মোট ১২টি ইউনিক শব্দ পাব। শব্দগুলো হলো:

'This', 'pasta', 'is', 'very', 'tasty', 'and', 'affordable', 'not', 'delicious', 'cheap', 'tastes', 'good'

এখন, শব্দগুলোর পুনরাবৃত্তি যাচাই করলে দেখা যাবে, 'pasta' শব্দটির পুনরাবৃত্তি হয়েছে মোট পাঁচবার, 'tasty' এবং 'affordable' শব্দদ্বয় উভয়ে দুইবার করে, 'delicious' একবার, 'good' একবার ইত্যাদি। লক্ষ করুন, 'pasta' শব্দটি সবচেয়ে বেশিবার প্রতীয়মান হয়েছে। অর্থাৎ রিভিউগুলো নিশ্চয়ই 'pasta' বিষয়ক। এবং 'pasta'র পাশাপাশি কিছু ইতিবাচক বা Positive শব্দেরও উপস্থিতি রয়েছে। যেমন, 'tasty', 'affordable' ইত্যাদি। এর দ্বারা রেস্টুরেন্ট ম্যানেজার বুঝে নেবে, 'pasta' এই মুহূর্তে সবচেয়ে জনপ্রিয় এবং এর চাহিদা বেশি। এভাবেই ব্যাগ অফ ওয়ার্ডস বিভিন্ন লেখার সম্পৃক্ততা, লেখার বিষয় শনাক্তকরণ ইত্যাদি কাজগুলো নিমিষেই করতে পারে। চিন্তা করুন, এতো গেল কেবল চারটি রিভিউ। বাস্তবে কাস্টমারদের হাজার এমনকি লাখো রিভিউ যাচাই করতে হয়। NLP এই ব্যাগ অফ ওয়ার্ডস টেকনিক বাস্তবায়নের মাধ্যমে অনায়াসেই অসংখ্য টেক্সট, লেখা বা ডকুমেন্ট প্রক্রিয়াকরণ করতে পারে। তবে ব্যাগ অফ ওয়ার্ডস এর কিছু জটিলতাও রয়েছে। বিশাল আকারের অসংখ্য ডকুমেন্ট প্রক্রিয়াকরণ যেকোনো কম্পিউটারের পক্ষেই গণনামূলকভাবে রিসোর্স ইন্টেন্সিভ। তাই খাটুনি কমাতে ব্যাগ অফ ওয়ার্ডস কিছু পন্থা অবলম্বন করে থাকে। ব্যাগ অফ ওয়ার্ডস শব্দাক্ষরের Case উপেক্ষা করে, অর্থাৎ 'pasta' এবং 'PaSta' কে একই বিবেচনা করে। এটি বিরাম চিহ্ন বা Punctuation Mark গুলো বাদ দিয়ে দেয়। কিছু শব্দ রয়েছে যাদের অন্তর্নিহিত অর্থ খুব একটা প্রভাব বিস্তার করে না, এদের **স্টপ ওয়ার্ডস** (Stop Words) বলা হয়। যেমন, 'is', 'the', 'of' ইত্যাদি। ব্যাগ অফ ওয়ার্ডস এই স্টপ ওয়ার্ডগুলোকে অপসারণ করে দেয়। শব্দে বানান ভুল থাকলে তা ঠিক করে নেয়। স্টেমিং এবং লেমাটাইজেশনের মাধ্যমে অনেকসময় এটি শব্দসমূহকে তাদের শব্দমূলে পরিণত করে। এভাবেই কোনো লেখার অর্থ বা প্রাসঙ্গিকতায় অবদান নেই এমন অসংখ্য শব্দ প্রক্রিয়াকরণের পূর্বেই বাতিল হয়ে যায়।

টিএফ-আইডিএফ (TF-IDF):

বিভিন্ন ডকুমেন্ট বা ফাইলে নির্দিষ্ট কিছু শব্দের গুরুত্ব বা বিরলতা (Rarity) নির্ধারণের পদ্ধতি হলো TF-IDF (Term Frequency–Inverse Document Frequency)। NLP মডেলগুলো গাণিতিক সূত্র ব্যবহার করে ডকুমেন্টে উপস্থিত প্রতিটি শব্দের একটি মূল্যমান নির্ণয় করে। শব্দের নির্দিষ্ট এই মূল্যমানই হলো ঐ শব্দের TF-IDF মান। TF-IDF এর প্রথম অংশ TF হলো Term Frequency যা মূলত কোনো ডকুমেন্টে একটি শব্দের পুনরাবৃত্তি সংখ্যা এবং ঐ ডকুমেন্টের মোট শব্দসংখ্যার অনুপাত (Ratio)। আর দ্বিতীয় অংশ IDF হলো Inverse Document Frequency যা মোট ডকুমেন্ট সংখ্যা এবং নির্দিষ্ট একটি শব্দ উপস্থিত রয়েছে এমন ডকুমেন্ট সংখ্যার অনুপাতের লগারিদম (Logarithm)। একটি শব্দের TF এবং IDF মানের গুণফলই হলো ঐ শব্দের TF-IDF মান। এই পদ্ধতিটি ব্যাগ অফ ওয়ার্ডসের অনেকটা উল্টো। ডকুমেন্টে যে শব্দগুলোর পুনরাবৃত্তি বেশি হয় সেই শব্দসমূহের TF-IDF মান হয় শূন্য। এর কারণ IDF একটি অনুপাতের লগারিদম। যখন বিভিন্ন ডকুমেন্টে একটি শব্দের অধিক পুনরাবৃত্তি হয় তখন অনুপাতটি একের কাছাকাছি পৌঁছায়। এবং একের লগারিদম সবসময় শূন্য। IDF এর মান শূন্য হওয়ার কারণে TFএর সাথে এর গুণফলও শূন্য হয়। অপরপক্ষে যে শব্দগুলো খুবই বিরল বা যার পুনরাবৃত্তি নেই বললেই চলে এমন সব শব্দের TF-IDF মান হয় শূন্যের বেশি। এর কারণও ঐ IDF মান। যখন কোনো শব্দের পুনরাবৃত্তি খুবই কম হয়, তখন IDF লগারিদমের মান হয় শূন্যের বেশি। যার কারণে TFএর সাথে এর গুণফলও হয়ে যায় শূন্যের বেশি। দুটি বাক্য দেখে নেওয়া যাক। একেকটি বাক্য একেকটি ডকুমেন্ট।

বাক্য-১: The car is driven on the road.

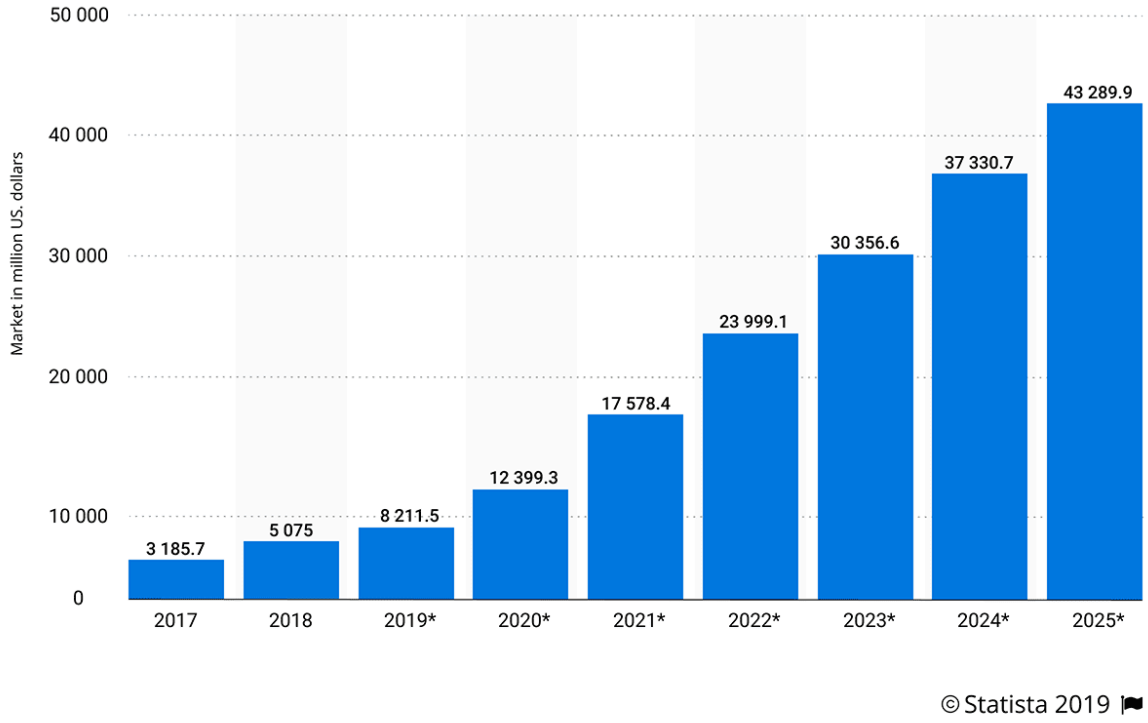
বাক্য-২: The truck is driven on the highway.

প্রথম বাক্যে ‘car’ এবং ‘road’ ব্যতিরেকে প্রত্যেকটি শব্দের TF-IDF মান শূন্য। আবার দ্বিতীয় বাক্যে ‘truck’ এবং ‘highway’ ছাড়া অন্য সকল শব্দের TF-IDF মান শূন্য। এর দ্বারা বোঝা গেল প্রথম বাক্যে ‘car’ এবং ‘road’ শব্দদ্বয় বিশেষ গুরুত্ব বহন করে। একইভাবে দ্বিতীয় বাক্যে ‘truck’ এবং ‘highway’ ছাড়া অন্যান্য শব্দগুলোর বিশেষ কোনো তাৎপর্য নেই। NLP মডেলগুলো শব্দের TF-IDF মান যাচাই করার মাধ্যমে বুঝতে পারে বিভিন্ন ডকুমেন্টে শব্দের প্রাচুর্যতা বা বিরলতা কেমন, কোন শব্দগুলো বিশেষ গুরুত্ব বহন করে।

NLP-র প্রয়োগ এবং ব্যবহার:

কৃত্রিম বুদ্ধিমত্তা চালিত মেশিন এবং মানুষের মিথস্ক্রিয়া নতুন কিছু নয়। মার্কেটিং এবং ব্যবসার উন্নতির স্বার্থে অসংখ্য কোম্পানি ডেটা সায়েন্স এবং মেশিন লার্নিং প্রযুক্তি প্রয়োগ করছে। বর্তমানে সবচেয়ে দ্রুত বর্ধনশীল AI প্রযুক্তিগুলোর মধ্যে NLP অন্যতম। NLP-র প্রয়োগ এবং প্রভাব সুদূরপ্রসারী। প্রতিনিয়তই বৃদ্ধি পাচ্ছে NLP-র চাহিদা এবং তৈরি

হচ্ছে NLP ব্যবহারের নতুন নতুন সব ক্ষেত্র। ২০১৯ Statista রিপোর্টে প্রকাশিত হয় যে ২০২৫ সাল নাগাদ NLP-র মার্কেট ৪৩.৯ বিলিয়ন ডলারে বৃদ্ধি পাবে। এখন জেনে নেওয়া যাক, NLP বহুমাত্রিক প্রয়োগ ও ব্যবহার সম্বন্ধে।



ব্যাংকিং-এ NLP:

ফাইন্যান্স (Finance) এবং ব্যাংকিং খাতে NLP ইতোমধ্যেই প্রতিষ্ঠা লাভ করেছে। NLP ব্যবহারের মাধ্যমে ঝুঁকি মূল্যায়ন (Risk Assessment) এবং পোর্টফোলিও অপ্টিমাইজেশান (Portfolio Optimization) থেকে আর্থিক অনুভূতি (Financial Sentiment) বিশ্লেষণ করা যায়। ইন্টারন্যাশনাল ব্যাংকার ম্যাগাজিনের মতে, ব্যাংকিং-এ NLP ব্যবহার বৃদ্ধির কারণ হল অধিকাংশ আর্থিক তথ্য লিখিত নথিতে (যেমন রিপোর্ট, প্রেস রিলিজ, অফিসিয়াল বিবৃতি) অন্তর্ভুক্ত করা হয়। ডিপ লার্নিং এবং NLP-র সংমিশ্রণ কম্পিউটারগুলিকে এই বিপুল সংখ্যক নথি পড়তে এবং অতি অল্প সময়ের মধ্যে প্রাসঙ্গিক তথ্য নির্বাচন করতে সক্ষম করে। ব্যাংকিং-এ NLP ব্যবহারের ক্ষেত্রে একটি অনন্য উদাহরণ হলো জেপি মরগ্যানের (JP Morgan) কন্ট্রাক্ট ইন্টেলিজেন্স (Contract Intelligence - COIN) সফটওয়্যার। এই সফটওয়্যার প্রচুর পরিমাণে আইনি নথি পর্যালোচনা করতে পারে। অনুমান করা হয় যে সফটওয়্যারটি ব্যবহারের কারণে জেপি মরগ্যানের আইনি দলের জন্য প্রতি বছর প্রায় ৩৬০,০০০ ঘন্টা সাশ্রয় হয়। ব্যাংকিং-এ NLP-র আবেদনের আরেকটি আশাপ্রদ ক্ষেত্র হল মার্কেট ট্রেন্ড (Market Trend) দেখা। নির্দিষ্ট কিছু সফটওয়্যার NLP ব্যবহার করে মাধ্যমে প্রতি সেকেন্ডে সোশ্যাল মিডিয়ায় প্রকাশিত হাজার হাজার মন্তব্য (Comment), অজস্র লাইক এবং শেয়ার, বিভিন্ন আদেশ উপদেশ, ইমোটিকন (Emoticon) এবং অসংখ্য স্ট্যাটাস

(Status) ও পোস্টের বিষয়গত অর্থ উদ্ধার করে এবং মানুষের স্বভাব, আবেগ, অনুভূতি, মন, মেজাজ ইত্যাদি বিশ্লেষণ করে। এই প্রক্রিয়াটিকে **সেন্টিমেন্ট অ্যানালাইসিস** (Sentiment Analysis) বলা হয়। সেন্টিমেন্ট অ্যানালাইসিসের মাধ্যমে ব্যাংক ম্যানেজাররা বিনিয়োগকারীদের পছন্দ অপছন্দ, মন মেজাজ, বিনিয়োগের প্রবণতা বুঝতে পারে যা যেকোনো আর্থিক সিদ্ধান্ত বিবেচনা করার জন্য একটি গুরুত্বপূর্ণ ফ্যাক্টর। ব্যাংকিং-এ সেন্টিমেন্ট অ্যানালাইসিসের একটি গুরুত্বপূর্ণ প্রয়োগ হল গ্রাহকরা সন্তুষ্ট কিনা তা বোঝা। ব্যাংকিং পরিষেবা সম্পর্কে গ্রাহকদের মন্তব্য বিশ্লেষণ করে, একটি ব্যাংক সহজেই জানতে পারে যে তারা সন্তুষ্ট কিনা।

স্বাস্থ্যসেবায় NLP:

এই মাত্র দেখেছি যে, ব্যাংকিং-এ যাবতীয় তথ্য বিভিন্ন ধরনের আর্থিক নথিতে ধারণ করা হয়, যার উপর NLP প্রয়োগ করা হয়। এই একই প্রক্রিয়া স্বাস্থ্যসেবায় বিদ্যমান। তবে এর নির্দিষ্ট একটি নাম রয়েছে, ক্লিনিকাল ডকুমেন্টেশন (Clinical Documentation)। আমেরিকান মেডিকেল অ্যাসোসিয়েশনের মতে, মার্কিন চিকিৎসকরা কাগজপত্র এবং প্রশাসনিক কাজে সপ্তাহে ১০ ঘন্টার বেশি সময় ব্যয় করেন। ফোর্বস (Forbes) রিপোর্ট করেছে যে একজন ডাক্তারের প্রতি ঘণ্টায় রোগীদের দেখার সময় গড়ে ২ ঘন্টা সময় নষ্ট হয় কেবল ফর্ম পূরণ করতে। NLP ডাক্তারদের প্রশাসনিক যন্ত্রণা উপশম করতে পারে। NLP **ভয়েস রিকগনিশনের** (Voice Recognition) মাধ্যমে একজন চিকিৎসক স্পিচ-টু-টেক্সট (Speech-to-Text) সফটওয়্যার ব্যবহার করে সহজেই তার কথা বা বক্তব্যকে লেখায় পরিণত করতে পারবে যা পরবর্তীতে একটি নথিথে সংরক্ষণ করা যায়। এটি কেবল সময়ই নয়, অর্থও বাঁচাতে পারে এবং হয়তো বাঁচাতে পারে মানুষের জীবন। অফুরান কাগজপত্র সংকলনের ক্লান্তিদায়ক সমস্যার সমাধান করা NLP-র একটি দিক। অন্য দিকটি হচ্ছে ক্লিনিকাল রেকর্ড থেকে মূল্যবান তথ্য পড়া এবং শনাক্ত করা। ইতোমধ্যেই ইলেক্ট্রনিক হেলথ রেকর্ড বা EHR ডাক্তারদের জীবনকে সহজ করে তুলেছে। এর সাথে NLP জুড়িয়ে দিলে দারুণ ফলাফল পাওয়া যায়। NLP সফটওয়্যারগুলো EHR পড়তে পারে এবং চোখের পলকে যে কোনও প্রাসঙ্গিক তথ্য বের করে আনে। স্বাস্থ্যসেবায় NLP-র অন্যান্য ব্যবহারের মধ্যে রয়েছে সুরক্ষিত স্বাস্থ্য তথ্য বা PHI সংরক্ষণ করা এবং রোগের লক্ষণগুলির ক্রস-রেফারেন্সিং করা। যেকোনো গুরুত্বপূর্ণ তথ্য সুরক্ষার স্বার্থে NLP দ্রুততার সাথে ডকুমেন্ট স্ক্যান করে সংবেদনশীল বা সেন্সেটিভ (Sensitive) তথ্যসমূহ শনাক্ত করে। এরপর সেই তথ্যগুলোকে তাদের সমতুল্য শব্দগুচ্ছ বা অস্পষ্ট তথ্য দ্বারা প্রতিস্থাপিত করে। এতে ডকুমেন্ট দুষ্কৃতিকারীর হাতে পরলেও সে কিছুই বুঝতে পারবে না। উপসর্গের ক্রস-রেফারেন্সিং-এর মাধ্যমে NLP আরও সূক্ষ্মতা ও সথিকতার সাথে রোগ নির্ণয় করতে পারে এবং রোগীর যথাযথ পর্যবেক্ষণ নিশ্চিত করে। এছাড়াও ক্লিনিকাল ডিসিশান সাপোর্ট, ক্লিনিকাল ট্রায়াল ম্যাচিং, এবং হেলথ প্রেডিকশানের মতো বহুক্ষেত্রেই NLP স্বাস্থ্যসেবায় অপরিহার্য হয়ে উঠছে।

শিক্ষায় NLP:

শিক্ষাক্ষেত্রে NLP-র প্রভাব অনস্বীকার্য। বাক্যের ব্যাকরণ সংশোধন করা NLP-র অনন্য ব্যবহারগুলোর একটি। কোনো রচনা বা প্যারাগ্রাফে ব্যাকরণগত ত্রুটি বা বানান ভুল আছে কিনা তা অনায়াসেই বলে দিতে পারে NLP। NLP অজস্র ডেটা স্ক্যানিং এবং প্রোসেসিং করে NLP এটাও বলে দিতে সক্ষম যে সঠিকভাবে লিখলে রচনাটি কেমন হতো। এতে একজন শিক্ষার্থী থেকে শুরু করে শিক্ষক, কর্পোরেট অফিসার, ম্যানেজারসহ অনেকেই ত্রুটিহীন এবং সাবলীল বক্তব্য বা লেখনি তৈরি করতে পারবে। এর একটি পরিচিত উদাহরণ হলো গ্র্যামারলি (Grammarly)। অনেকে লেখার সৌন্দর্য এবং শব্দচয়নের মাধুর্য ঠিক রাখতে গ্র্যামারলি প্রিমিয়াম সাবস্ক্রিপশান নিয়ে থাকেন। NLP-র আরেকটি গুরুত্বপূর্ণ ব্যবহার হলো লেখার সারমর্ম উদ্ধার করা যাকে **টেক্সট সামারাইজেশন** (Text Summarization) বলা হয়। প্রায়শই, কোনো প্রবন্ধ জমা দিতে গেলে বা একটি নিবন্ধ লিখতে গেলে শব্দ দৈর্ঘ্যের সীমা অতিক্রম হয়ে যায়। টেক্সট সামারাইজেশন হলো সুদীর্ঘ একটি লেখার অর্থ এবং মূল বিষয় ঠিক রেখে নির্দিষ্ট একটি দৈর্ঘ্যে সংকুচিত করে সংক্ষিপ্ত করা। হাতেকলমে লেখা সংক্ষিপ্তকরণ খুবই সময়সাপেক্ষ এবং ক্লান্তিকর একটি কাজ। কিন্তু আজ NLP-র সহায়তায় স্বয়ংক্রিয়ভাবে চোখের পলকেই লেখা সংকুচিত হয়ে যাচ্ছে, এবং এতে অর্থ বা প্রাসঙ্গিকতার পরিবর্তন হচ্ছে না। বিশাল কোনো প্রবন্ধ পড়ার সময় না থাকলে টেক্সট সামারাইজেশনের মাধ্যমে লেখার মূল বিষয়বস্তু নিমিষেই বুঝে নেওয়া সম্ভব। টেক্সট সামারাইজেশন আবার দুই প্রকার। একটি হলো এক্সট্র্যাক্টিভ (Extractive) সামারাইজেশন যা সরাসরি ইনপুট টেক্সট নিয়ে সেই ইনপুট টেক্সটের অপ্রয়োজনীয় অংশ বাদ দিয়ে শুধু অর্থ ধারণ করে এমন অংশটুকু রেখে দেয়। আরেকটি হলো অ্যাবস্ট্রাক্টিভ (Abstractive) সামারাইজেশন যা ইনপুট টেক্সট সরাসরি ব্যবহার না করে ডিপ লার্নিং ব্যবহার করে প্রথমে তার সারমর্ম উদ্ধার করে। তারপর সেই সারমর্মের ভিত্তিতে নতুন এবং সংক্ষিপ্ত একটি আউটপুট টেক্সট তৈরি করে। এই প্রক্রিয়াটি নিশ্চয়ই প্রথমটি থেকে উন্নত ফলাফল দেবে, তবে এটি তুলনামূলকভাবে জটিল প্রক্রিয়া।

উৎপাদন শিল্পে NLP:

উৎপাদন শিল্পে NLP ব্যবহারের প্রধান দুটি উদাহরণ হলো প্রকিউরমেন্ট (Procurement) এবং লজিস্টিকস (Logistics)। সাপ্লাই চেইনের যেকোনো পর্যায়ে লক্ষ লক্ষ বিল, ইনভয়েস, ডেলিভারি নোট এবং অন্যান্য অনুরূপ নথিগুলি নিয়মিত পরিচালনা করে থাকে অসংখ্য ইন্ডাস্ট্রি। NLP সফটওয়্যারগুলো যাবতীয় নথি স্ক্যানিং, শ্রেণীবদ্ধকরণ, লেবেলিং করার মাধ্যমে পুরো সরবরাহ প্রক্রিয়াকে সুগমকরণে সাহায্য করতে পারে। স্বয়ংক্রিয় ইনভেন্টরি ম্যানেজমেন্ট বাস্তবায়নের মাধ্যমে স্বল্প সময়েই প্রাসঙ্গিক ডেটা বের করে এনে সরাসরি অ্যাকাউন্টিং ডকুমেন্টেশনে সংরক্ষণ করা যেতে পারে। এতে বেঁচে যাবে ইন্ডাস্ট্রিগুলোর সময় এবং অর্থ। NLP-র আরেকটি গুরুত্বপূর্ণ ব্যবহারের ক্ষেত্র হলো ওয়েব স্ক্র্যাপিং (Web Scraping)। এর অর্থ হলো পরিবহণের হার, জ্বালানীর হার এবং অন্যান্য বেঞ্চমার্ক (Benchmark) অনুসন্ধান করে এবং তুলনা করার মাধ্যমে খরচের ক্ষেত্র শনাক্তকরণ এবং খরচ-সঞ্চয় করার

সর্বোত্তম পন্থা নির্ধারণ করা। এরপর আসে NLP সেন্সর (Sensor)। সেন্সরের কাজ সেন্স (Sense) করা। সাধারণ সেন্সরগুলোয় কৃত্রিম বুদ্ধিমত্তা, ডিপ লার্নিং এবং NLP সংযুক্ত করে উন্নত উৎপাদন প্রক্রিয়া সম্পাদন করা যায়। একটি NLP সেন্সর শুধুমাত্র শুনতে এবং দেখতে পারে না, এটি কথোপকথন বা ভোকাল ইনপুটও বুঝতে পারে। যদি কেউ পাওয়ারপ্ল্যান্টের নিষিদ্ধ একটি এলাকায় প্রবেশ করে, তাহলে NLP সেন্সর শব্দ, কথোপকথন এবং অন্যান্য ইনপুট রেকর্ড করে বিপদের মাত্রা অনুযায়ী কেমন প্রতিরোধ ব্যবস্থা নিতে হবে তা নির্ধারণ করতে পারে। উৎপাদন প্রক্রিয়া স্বয়ংক্রিয়করণে NLP অন্যান্য প্রযুক্তির সাথে একীভূত করা যেতে পারে। স্বয়ংক্রিয় উৎপাদনে NLP ব্যবহারের একটি উদাহরণ হলো, উৎপাদনে ব্যবহৃত সরঞ্জামগুলি ১০০ শতাংশ কর্মদক্ষ কিনা তা NLP শনাক্ত করতে সক্ষম। এতে উৎপাদনে অর্থের সঞ্চয় হয় এবং সামগ্রিক উৎপাদন ক্ষমতা বৃদ্ধি পায়।

মার্কেটিং এবং রিটেইলে NLP:

কাস্টমার সার্ভিসের স্বয়ংক্রিয়করণ বোধ হয় NLP প্রয়োগের সবচেয়ে উৎকৃষ্ট উদাহরণগুলোর একটি। কৃত্রিম ভয়েস এবং কৃত্রিম বুদ্ধিমত্তা সম্বলিত **চ্যাটবট** (Chatbot) আবির্ভাবের মধ্য দিয়ে উদ্ভাবিত হয়েছে ভার্চুয়াল অ্যাসিস্টেন্ট (Virtual Assistant) যা প্রতিনিয়তই পরিষেবার স্বয়ংক্রিয়করণে অবদান রাখছে। কৃত্রিম বুদ্ধিমত্তা সম্বলিত চ্যাটবটগুলো কাস্টমারদের কোম্পানির হেল্প ডেস্কের নির্দিষ্ট কোনো প্রতিনিধির কাছে রাউট (Route) করে দেয় কিংবা চ্যাটবটগুলো নিজেরাই প্রতিনিধির ভূমিকা পালন করে কাস্টমারদের সাথে দিব্যি কথোপকথন চালিয়ে যায় এবং সাধারণ প্রশ্নগুলোর উপযুক্ত উত্তর প্রদান করে। এতে কোম্পানিগুলোর আলাদাভাবে FAQ (Frequently Asked Questions) পেজ তৈরি করার প্রয়োজন হবে না। তবে জটিল প্রশ্ন বা খুবই নির্দিষ্ট কোনো জিজ্ঞাসা থাকলে মানুষই চ্যাটবট অপেক্ষা সামঞ্জস্যপূর্ণ উত্তর দেয়। রিটেইল ব্যবসায় NLP নিদারুণ সফলতা অর্জন করেছে। মার্কিন বিজনেস ম্যাগাজিন ফরচুনের (Fortune) মতে, কাস্টমার এক্সপেরিয়েন্স (Customer Experience) নিশ্চিত করার লক্ষ্যে NLP-র চাহিদা বহুগুণে বৃদ্ধি পেয়েছে। মার্কেটিং-এ NLP-র গুরুত্ব অপরিসীম। ভোক্তা, গ্রাহক বা কাস্টমার সংক্রান্ত ডেটা প্রক্রিয়াকরণ এবং কাস্টমার প্রোফাইলিং করে NLP। বিভিন্ন ফোরাম (Forum), ওয়েবসাইট কিংবা সোশ্যাল মিডিয়া থেকে একটি প্রোডাক্ট বা সার্ভিস সম্পর্কে কাস্টমারদের অভিমত বা ফিডব্যাক (Feedback) প্রক্রিয়াকরণ করার মাধ্যমে NLP বুঝতে পারে প্রোডাক্ট বা সার্ভিসের চাহিদা কেমন এবং কোন কাস্টমার কিধরণের সামগ্রী পছন্দ করে। এর ভিত্তিতে কোম্পানিগুলো ব্যাপক প্রচারণা চালায় এবং বিজ্ঞাপন ছড়িয়ে দেয়। কাস্টমারদের পছন্দসই দ্রব্যসমূহের যথাযথ মার্কেটিং হওয়ায় বিক্রয় বহুগুণে বেড়ে যায়। ফরচুনের প্রকাশিত তথ্য অনুযায়ী, বিশ্বব্যাপী NLP মার্কেট শেয়ারে দ্বিতীয় অবস্থানে রয়েছে রিটেইল এবং মার্কেটিং, যা মোটেও আশ্চর্যজনক নয়। বর্তমান বিশ্বে অসংখ্য কোম্পানি তাদের মার্কেটিং এবং রিটেইল ব্যবসায় NLP ব্যবহার করছে যার মধ্যে গুগল, অ্যামাজন, মাইক্রোসফট, আইবিএম, হেউলেট প্যাকার্ড (Hewlett Packard), ইন্টেলের মতো নামিদামি সব ব্র্যান্ড অন্তর্ভুক্ত রয়েছে।

তথ্য প্রযুক্তিতে NLP:

সাইবার নিরাপত্তার (Cybersecurity) দুনিয়ায় একটি সিস্টেমকে সুরক্ষিত করতে বিভিন্ন ক্ষতিকর **স্প্যাম (Spam)** মন্তব্য, স্প্যাম ইমেইল, অপ্রয়োজনীয় মেসেজ, বোগাস (Bogus) রিকুয়েস্ট ইত্যাদি প্রতিহত করতে হয়। NLP দ্বারা এই কাজটি বেশ লাভজনকভাবে করা সম্ভব। এক্ষেত্রেও NLP-র ব্যবহার আগের মতোই। অজস্র ইমেইল, রিকুয়েস্ট, কমেন্ট স্ক্যান করে NLP তার অর্থ উদ্ধার করে এবং সিদ্ধান্ত নেয় মেসেজগুলো স্প্যাম কিনা। স্প্যাম হলে মেসেজগুলো অপসারণ করে। সাইবার নিরাপত্তায় NLP-র আরেকটি উল্লেখযোগ্য ব্যবহার হলো **ডেটা এক্সফিল্ট্রেশন** (Data Exfiltration) প্রতিরোধ। বিভিন্ন ক্ষতিকর ম্যালওয়্যারের (Malware) মাধ্যমে কোনো সিস্টেম বা ডাটাবেজ থেকে অননুমোদিতভাবে সংবেদনশীল তথ্য যেমন ইউজার পাসওয়ার্ড হাতিয়ে নেওয়ার প্রচেষ্টাই হলো ডেটা এক্সফিল্ট্রেশন। এটা এক ধরণের ডেটা ব্রিচ (Data Breach)। NLP ক্ষতিকর ডোমেইন (Domain), আইপি (IP) বা ফিশিং ইমেইল (Phishing Email) খুঁজে বের করে সেগুলোকে ব্লক করে দেয়। এতে তারা কোনো ধরণের ম্যালওয়্যার, ভাইরাস কিংবা স্প্যাম ইমেইল পাঠাতে পারে না। এফবিআই (FBI) রিপোর্টে উঠে আসে, ২০২০ সালে ফিশিং আক্রমণ খুবই প্রচলিত ছিল। তাই তথ্য প্রযুক্তির জগতেও NLP গুরুত্বপূর্ণ অবদান রাখতে পারে।

NLP-র ভবিষ্যৎ:

প্রাকৃতিক ভাষা প্রক্রিয়াকরয়াণে NLP যদিও বেশ আশাপ্রদ ফলাফল প্রদর্শন করেছে এখনও বহুদূর পথ পাড়ি দেয়া বাকি। ২০১৬ সালের মার্চ মাসে NLP-র পরীক্ষামূলক প্রোজেক্ট হিসেবে টে (Tay) নামক কৃত্রিম বুদ্ধিমত্তা সম্পন্ন একটি চ্যাটবট তৈরি করে মাইক্রোসফট। টে-এর প্রোজেক্টটি টুইটারে প্রকাশিত হয় এবং এর লক্ষ্য ছিল মানুষের সাথে কথোপকথনের মাধ্যমে এটি আরও বুদ্ধিমান এবং চমকপ্রদ সব টুইট বার্তা জেনারেট করবে। অথচ এটি প্রকাশের ১৬ ঘণ্টার মাথায় বর্ণবাদী এবং আপত্তিকর সব টুইট করে বসে যার কারণে টে-কে দ্রুত এই প্ল্যাটফর্ম থেকে সরিয়ে নেওয়া হয়েছিল। এই তিক্ত অভিজ্ঞতার পর মাইক্রোসফট তাদের ভুল সংশোধন করে এবং আরও উন্নত বুদ্ধিসম্পন্ন চ্যাটবট জো (Zo) কে নিয়ে আসে। জো মানুষের কথোপকথন শনাক্ত করতে এবং নতুন কথোপকথন তৈরি করতে সক্ষম ছিল। NLP-র ভবিষ্যৎ যদিও বেশ চ্যালেঞ্জিং বলে প্রতীয়মান হয়, এর গবেষণায় প্রচুর পরিমাণে বিনিয়োগ করা হচ্ছে। প্রতিনিয়তই বিভিন্ন জার্নালে পাবলিশ হচ্ছে অসংখ্য প্রবন্ধ এবং নিবন্ধ যা NLP-র ব্যাপক উৎকর্ষ সাধন করছে। বিগত বছরগুলোর তুলনায় NLP দ্রুতগতিতে বিকশিত হচ্ছে। অদূর ভবিষ্যতে হয়তো আমরা এমন সফটওয়্যার অ্যাপ্লিকেশন ব্যবহার করবো যা আমাদের জীবনযাত্রার মানকে করবে আরও উন্নত এবং আরামদায়ক।

রেফারেন্স

1. https://www.youtube.com/watch?v=CMrHM8a3hqw
2. https://www.youtube.com/watch?v=fLvJ8VdHLA0
3. https://www.youtube.com/watch?v=fOvTtapxa9c
4. https://www.investopedia.com/terms/n/natural-language-processing-nlp.asp
5. https://towardsdatascience.com/a-detailed-novice-introduction-to-natural-language-processing-nlp-90b7be1b7e54
6. https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63
7. https://medium.datadriveninvestor.com/the-brief-history-of-nlp-c90f331b6ad7
8. https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce
9. https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37
10. https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696
11. https://medium.com/@jeevanchavan143/nlp-tokenization-stemming-lemmatization-bag-of-words-tf-idf-pos-7650f83c60be
12. https://medium.com/mlearning-ai/nlp-tokenization-stemming-lemmatization-and-part-of-speech-tagging-9088ac068768
13. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/#:~:text=A%202%2Dgram%20(or%20bigram,or%20%E2%80%9Con%20Analytics%20Vidhya%E2%80%9D.
14. https://medium.com/@raghvendra.zarkar18/natural-language-processing-65f82c8dd7e0
15. https://towardsdev.com/lemmatization-in-natural-language-processing-nlp-5d2434527766

16. https://towardsdatascience.com/stemming-vs-lemmatization-in-nlp-dea008600a0#:~:text=Stemming%20and%20Lemmatization%20are%20algorithms,same%20word%20in%20different%20tenses.
17. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
18. https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/
19. https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22
20. https://www.ideta.io/blog-posts-english/nlp-use-cases
21. https://mobidev.biz/blog/natural-language-processing-nlp-use-cases-business
22. https://www.analyticsvidhya.com/blog/2021/05/interesting-nlp-use-cases-every-data-science-enthusiast-should-know/
23. https://datasaur.ai/blog-posts/important-use-cases-of-nlp
24. https://www.kdnuggets.com/2019/05/guide-natural-language-processing-nlp.html