

தமிழ் மொழியில் இயற்கை மொழி செயலாக்கத்தின் (NLP) பயன்பாடுகளும் சவால்களும்

முனைவர்.நா.பிரபா, உதவி பேராசிரியர், வணிகவியல் மற்றும் கணினி பண்பாட்டியல் துறை, மன்னர் திருமலை
நாயக்கர் கல்லூரி, மதுரை, தமிழ்நாடு, இந்தியா.
ORCID: <https://orcid.org/0009-0003-6818-7801>

அறிமுகம்

இயற்கை மொழி செயலாக்கம் (Natural Language Processing - NLP) என்பது கணினி அறிவியலின் ஒரு முக்கியமான துணைத்துறையாகும். இது கணினிகள் மனித மொழியைப் புரிந்துகொள்ளவும், விளக்கவும், கையாளவும், மற்றும் உருவாக்கவும் உதவுகின்ற வழிமுறைகள் மற்றும் தொழில்நுட்பங்களை உள்ளடக்கியது. NLP-யின் அடிப்படை நோக்கம், மனித மொழிக்கும் கணினி மொழிக்கும் இடையே ஒரு பாலத்தை உருவாக்குவதாகும். இந்தத் தொழில்நுட்பம், தற்போது உலகின் பல முக்கிய மொழிகளில் (ஆங்கிலம், சீனம், ஸ்பானியம், ஜெர்மன் போன்றவை) மிகவும் வளர்ச்சி பெற்றுள்ளது. ஆனால், தமிழ் மொழியைப் பொறுத்தவரை, NLP துறை இன்னும் முதிர்ச்சியடையாத ஒரு நிலையிலேயே உள்ளது. தமிழ், உலகின் தொன்மையான மொழிகளுள் ஒன்றாக இருந்த போதிலும், அதன் தனித்துவமான இலக்கணக் கட்டமைப்பு, வேற்றுமை உருபுகளின் சிக்கலான அமைப்பு, உயிர்-மெய் எழுத்துக்களின் இணைப்பு முறை, மற்றும் இருமொழி (Diglossia) நிலை ஆகியவை NLP-யின் வளர்ச்சிக்குப் பெரும் சவால்களாக அமைந்துள்ளன. இருப்பினும், சமீபத்திய ஆண்டுகளில், தமிழ் மொழிக்கான NLP கருவிகளை உருவாக்குவதில் பல முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இக்கட்டுரை, தமிழ் மொழியில் NLP-யின் பயன்பாடுகள், அதன் கல்வியியல் முக்கியத்துவம், தற்போதைய சவால்கள், மற்றும் எதிர்கால வாய்ப்புகளை ஒரு முழுமையான கல்வியியல் மதிப்பீட்டிற்கு உட்படுத்துகிறது.

NLP-யின் அடிப்படைத் தத்துவங்களும் தமிழ் மொழியின் தனித்தன்மையும்

NLP-யின் அடிப்படைச் செயல்பாடுகள், மொழியியல் பகுப்பாய்வின் பல்வேறு நிலைகளை உள்ளடக்கியவை. அவை:

- **எழுத்து மற்றும் சொல் மட்டம் (Lexical Level):** சொற்களின் அடிப்படை வடிவங்களைக் கண்டறிதல் (Tokenization), வேர்ச் சொற்களைப் பிரித்தல் (Lemmatization/Stemming).
- **இலக்கண மட்டம் (Syntactic Level):** வாக்கியத்தில் உள்ள சொற்களுக்கிடையேயான இலக்கண உறவுகளைப் பகுப்பாய்வு செய்தல் (Parsing).
- **சொற்பொருள் மட்டம் (Semantic Level):** சொற்கள் மற்றும் வாக்கியங்களின் பொருளைப் புரிந்துகொள்ளுதல் (Word Sense Disambiguation).
- **செயல்முறை மட்டம் (Pragmatic Level):** சூழ்நிலைக்கேற்ப மொழியின் பயன்பாட்டைப் புரிந்துகொள்ளுதல்.

தமிழ் மொழி, இந்த ஒவ்வொரு மட்டத்திலும் ஆங்கிலம் போன்ற இந்தோ-ஐரோப்பிய மொழிகளிலிருந்து முற்றிலும் வேறுபட்ட அமைப்புகளைக் கொண்டுள்ளது. எடுத்துக்காட்டாக, தமிழில் சொற்கள் ஒன்றோடொன்று ஒட்டிக்கொள்ளும் (Agglutinative) தன்மை கொண்டவை. ஒரு தமிழ்ச் சொல், அதன் வேர், பல்வேறு உருபுகள், மற்றும் இடைநிலைகள் ஆகியவற்றை ஒன்று சேர்த்து ஒரு நீண்ட சொல்லை உருவாக்கும். ஆங்கிலத்தில் ஒரு முழு வாக்கியத்தை வெளிப்படுத்துவதற்கு ஐந்து அல்லது ஆறு சொற்கள் தேவைப்படும், அதே கருத்தைத் தமிழில் ஒரு சில சொற்களிலேயே வெளிப்படுத்த இயலும். இந்த ஒட்டுக்கொள்ளும் (Agglutinative)

தன்மையை NLP கருவிகள் துல்லியமாகப் பகுப்பாய்வு செய்வது என்பது மிகவும் சிக்கலான ஒரு கணினியியல் பிரச்சனையாகும். மேலும், தமிழில் "எழுவாய்-பயனிலை-கருப்பொருள்" (SOV) என்ற சொற்களஞ்சிய அமைப்பு நிலவுகிறது. ஆனால், இலக்கண விதிகளின் அடிப்படையில், சொற்களின் வரிசை மாற்றம் கூடியதாகும். இந்த நெகிழ்வுத்தன்மை (Free Word Order) NLP-யின் சோதனை வழிமுறைகளுக்கு (Parsing Algorithms) மிகவும் கடினமான ஒரு சவாலாக அமைகிறது. இந்தத் தனித்தன்மைகளைப் புரிந்துகொள்வதே தமிழ் NLP-யின் முதல் படியாகும்.

தமிழ் NLP-யின் முக்கிய பயன்பாடுகள்

இயந்திர மொழிபெயர்ப்பு (Machine Translation)

தமிழ் NLP-யின் மிகவும் முக்கியமான பயன்பாடு இயந்திர மொழிபெயர்ப்பு ஆகும். தமிழ் பேசும் மக்கள் உலகம் முழுவதும் பரவியுள்ளதால், தமிழிலிருந்து ஆங்கிலம், இந்தி, மலையாளம், தெலுங்கு போன்ற மொழிகளுக்கும், அவற்றிலிருந்து தமிழுக்கும் மொழிபெயர்க்கும் திறன் மிகவும் அவசியமானதாகும். நரம்பு மொழிபெயர்ப்பு மாதிரிகள் (Neural Machine Translation - NMT) தற்போது தமிழ்-ஆங்கிலம் மொழிபெயர்ப்பில் 60-70% வரை துல்லியத்தை அடைந்துள்ளன. இது ஒரு நாளைக்கு ஆயிரக்கணக்கான ஆவணங்களை மொழிபெயர்க்க உதவுகிறது. கல்வி, வணிகம், மருத்துவம், மற்றும் சட்டத் துறைகளில் இந்த மொழிபெயர்ப்பு கருவிகள் பெரும் பயனை வழங்குகின்றன.

பேச்சு அங்கீகாரமும் பேச்சிலிருந்து உரைக்கும் மாற்றமும் (Speech-to-Text)

தமிழ் பேச்சு அங்கீகாரம் (Automatic Speech Recognition - ASR) என்பது மற்றொரு முக்கியமான பயன்பாடாகும். தமிழில் பல்வேறு பேச்சு வழக்குகள் (Dialects) உள்ளன – சென்னைத் தமிழ், ஈழத் தமிழ், மலேசியத் தமிழ், குமரி தமிழ் என்று ஒவ்வொரு பிராந்தியத்திலும் பேச்சு வழக்குகள் வேறுபடுகின்றன. AI-யின் இயந்திரக் கற்றல் முறைமைகள், இந்த வேறுபாடுகளை இனங்கண்டு, பேச்சைத் துல்லியமாக எழுத்து வடிவில் மாற்றும் திறனை வளர்த்துள்ளன. இந்தத் தொழில்நுட்பம், ஒலிப்பதிவுகளை உரை வடிவில் மாற்றுதல், தமிழ் திரைப்படங்களின் வசனங்களைத் தானியங்கி முறையில் படியெடுத்தல், மற்றும் பார்வையற்றோருக்கான தமிழ் உரை வாசிப்புக் கருவிகள் ஆகியவற்றில் பயன்படுகின்றன.

உரைச் சுருக்கம் மற்றும் தகவல் பிரித்தெடுத்தல் (Text Summarization & Information Extraction)

தமிழில் நாள்தோறும் பல ஆயிரம் செய்திகள், கட்டுரைகள், ஆய்வுக் கட்டுரைகள் வெளியாகின்றன. இவற்றைச் சுருக்கமாகவும், பயனுள்ள தகவல்களைப் பிரித்தெடுக்கவும் NLP கருவிகள் பயன்படுகின்றன. எடுத்துக்காட்டாக, ஒரு நீண்ட தமிழ்ச் செய்தியை, அதன் முக்கிய நிகழ்வுகளை மட்டும் தேர்ந்தெடுத்து, மூன்று அல்லது நான்கு வாக்கியங்களில் சுருக்கித் தருவது ஒரு முக்கியமான பயன்பாடாகும். மேலும், தமிழ் உரைகளிலிருந்து பெயர்கள், இடங்கள், காலம், பொருட்கள் ஆகியவற்றைப் பிரித்தெடுப்பது (Named Entity Recognition - NER) அரசு ஆவணங்கள், சட்ட ஆவணங்கள், மற்றும் வணிகத் தரவுகளைப் பகுப்பாய்வு செய்ய உதவுகிறது.

சொற்பொருள் பகுப்பாய்வு மற்றும் உணர்வு பகுப்பாய்வு (Sentiment Analysis)

தமிழ் சமூக ஊடகங்களில் (Twitter, Facebook, YouTube) ஒவ்வொரு நாளும் இலட்சக்கணக்கான கருத்துகள் பதிவாகின்றன. இவற்றில் உள்ள நேர்மறை, எதிர்மறை, மற்றும் நடுநிலை உணர்வுகளைப் பகுப்பாய்வு செய்ய NLP பயன்படுகிறது. இது அரசியல் கணிப்புகள், சந்தை ஆராய்ச்சி, தயாரிப்பு மதிப்பீடுகள், மற்றும் பொது கருத்துக் கணிப்புகளில் மிகவும் பயனுள்ளதாக உள்ளது. தமிழில் உள்ள பழமொழிகள், உவமைகள், மற்றும்

முரண்பாடுகள் (Irony/Sarcasm) ஆகியவற்றைச் சரியாகப் பகுப்பாய்வு செய்வதே இங்கு மிகப்பெரிய சவாலாக உள்ளது.

தமிழ் தேடுபொறிகள் மற்றும் பரிந்துரை அமைப்புகள்

தமிழில் திறமையான தேடுபொறிகளை உருவாக்குவதற்கு NLP மிகவும் இன்றியமையாதது. ஒரு பயனர் "பாரதி கவிதை" என்று தேடும்போது, அதற்கு சரியான கவிதைகளை மட்டுமல்லாமல், பாரதியின் வாழ்க்கை வரலாறு, அவரது பாணி, மற்றும் தொடர்புடைய பிற கவிதைகளையும் பரிந்துரைக்கும் அமைப்புகள் NLP-யின் அடிப்படையில் இயங்குகின்றன. மேலும், தமிழ் இலக்கியப் படைப்புகளுக்கான பரிந்துரை அமைப்புகள், ஒரு வாசகரின் வாசிப்பு வரலாற்றைப் பகுப்பாய்வு செய்து, அவருக்கு மிகவும் பொருத்தமான அடுத்த நூலைப் பரிந்துரைக்கும் திறன் கொண்டவை.

தமிழ் NLP-யின் முக்கிய சவால்கள்

தரவுப் பற்றாக்குறை (Data Scarcity)

தமிழ் NLP-யின் மிகப்பெரிய சவால் தரவுப் பற்றாக்குறை ஆகும். ஆங்கில மொழிக்கு இலட்சக்கணக்கான கற்பித்தல் தரவுத் தொகுப்புகள் (Training Datasets) கிடைத்தாலும், தமிழுக்கு அப்படிப்பட்ட பெரிய அளவிலான தரவுத்தளங்கள் இல்லை. இயந்திரக் கற்றல் மாதிரிகள், குறிப்பாக ஆழ்ந்த கற்றல் (Deep Learning) மாதிரிகள், பயனுள்ள வெளியீட்டைத் தருவதற்கு ஏராளமான தரவுகள் தேவைப்படுகின்றன. தமிழின் பல்வேறு வட்டார வழக்குகள், இலக்கிய நடைகள், பேச்சு வழக்குகள் ஆகியவற்றை உள்ளடக்கிய போதுமான தரவுத் தொகுப்புகள் இல்லாததால், NLP கருவிகளின் துல்லியம் குறைந்துவிடுகிறது.

ஒட்டுக்கொள்ளும் மொழியியல் அமைப்பு (Agglutinative Morphology)

முன்னர் குறிப்பிட்டபடி, தமிழ் ஒரு ஒட்டுக்கொள்ளும் மொழியாகும். ஒரு சொல் பல உருபுகளை ஒட்டிக்கொண்டு, மிக நீண்டதாக மாறும். இந்தச் சொற்களைக் கையாள்வதற்கு, NLP கருவிகள் மிகச் சிக்கலான உருபனியல் பகுப்பாய்விகள் (Morphological Analyzers) தேவைப்படுகின்றன. தற்போது இருக்கும் தமிழ் உருபனியல் பகுப்பாய்விகள், பெரும்பாலும் சொற்களஞ்சியம் சார்ந்த (Lexicon-based) முறைகளைப் பயன்படுத்துகின்றன. இவை புதிய சொற்களைச் சரியாகப் பகுப்பாய்வு செய்யத் தவறுகின்றன.

பன்மொழி மற்றும் இருமொழி நிலை (Diglossia)

தமிழ் மொழியில் பேச்சுத் தமிழும், எழுத்துத் தமிழும் (செந்தமிழ்) வேறுபட்டவை. பேச்சுத் தமிழில் பல சொற்கள் மாற்றப்பட்டும், ஒலிப்பு வேறுபாடுகளுடனும் பயன்படுத்தப்படுகின்றன. NLP கருவிகள் பெரும்பாலும் எழுத்துத் தமிழில் பயிற்சியளிக்கப்படுவதால், பேச்சுத் தமிழைச் சரியாகப் பகுப்பாய்வு செய்வதில் சிக்கல் ஏற்படுகிறது. மேலும், சமூக ஊடகங்களில் பயனர்கள் தமிழை ஆங்கில எழுத்துகளில் எழுதும் (Tanglish) வழக்கமும் உள்ளது. இந்த மாற்றங்களைச் சமாளிப்பது NLP-க்கு ஒரு கடினமான பணியாகும்.

தரப்படுத்தப்பட்ட குறியீட்டு முறைகள் இல்லாமை

தமிழ் எழுத்துகளை யுனிகோட் (Unicode) முறையில் குறியீடு செய்வது பொதுவானதாக இருந்தாலும், பழைய தமிழ் எழுத்துருக்கள் (கிரந்தம், வட்டெழுத்து) மற்றும் சில சிறப்பு எழுத்துகள் (எடுத்துக்காட்டாக, ஸ்ரீ, ஹ, ஷ போன்றவை) யுனிகோட்டில் சரியான முறையில் குறியீடு செய்யப்படாத நிலை உள்ளது. இது பழைய இலக்கிய ஆவணங்களை டிஜிட்டல் மயமாக்குவதில் பெரும் சிக்கலை உருவாக்குகிறது.

தமிழ் NLP-யின் எதிர்காலச் சாத்தியங்களும் தீர்வுகளும்

இந்தச் சவால்கள் இருந்த போதிலும், தமிழ் NLP துறையில் பல நம்பிக்கையான முன்னேற்றங்கள் ஏற்பட்டுள்ளன.

- **கூட்டுத் தரவுத் தொகுப்பு உருவாக்கம்:** தமிழ்த் தரவுத் தொகுப்புகளை உருவாக்குவதற்குப் பல கல்வி நிறுவனங்களும், தொழில்நுட்ப நிறுவனங்களும் ஒன்றிணைந்து, திறந்த மூல (Open Source) தரவுத் தொகுப்புகளை உருவாக்கி வருகின்றன. எடுத்துக்காட்டாக, "தமிழ் NLP சமூகம்" மூலம் சுமார் 10 லட்சம் தமிழ் வாக்கியங்கள் கொண்ட ஒரு தரவுத் தொகுப்பு உருவாக்கப்பட்டுள்ளது.
- **முன்பயிற்சி பெற்ற மொழி மாதிரிகள் (Pre-trained Language Models):** பெரிய மொழி மாதிரிகள் (Large Language Models) தமிழ் மொழியிலும் பயிற்சியளிக்கப்படுகின்றன. குறிப்பாக, "பாரதி" என்ற பெயரில் ஒரு தமிழ் மொழி மாதிரி உருவாக்கப்பட்டு, அது கவிதை எழுதுதல், கதை உருவாக்குதல், மொழிபெயர்ப்பு ஆகியவற்றில் நல்ல முடிவுகளைத் தருகின்றது.
- **பேச்சு அங்கீகாரத்தில் புதிய மேம்பாடுகள்:** தமிழ் பேச்சு அங்கீகாரத்தில், ஆழ்ந்த நரம்பு வலையமைப்புகள் (Deep Neural Networks) மூலம் 80% வரை துல்லியம் அடையப்பட்டுள்ளது. வட்டார வழக்குகளைப் புரிந்துகொள்ளும் திறனும் மேம்பட்டு வருகிறது.
- **நெறிமுறை கட்டமைப்புகள்:** தமிழ் NLP-யின் தரவுகளைப் பயன்படுத்தும் போது, தனியுரிமை, கலாச்சார உணர்வு, மற்றும் பாகுபாடு தவிர்த்தல் போன்ற நெறிமுறைக் கட்டமைப்புகளை உருவாக்குவதில் ஆய்வாளர்கள் கவனம் செலுத்துகின்றனர்.
- **கல்வியில் தமிழ் NLP-யின் பங்கு**
- தமிழ் NLP-யின் மிகப்பெரிய பயன்பாடு கல்வித் துறையில்தான். தமிழ் மொழியை முதல் மொழியாகக் கொண்ட பள்ளி மாணவர்களுக்கும், தமிழை இரண்டாம் மொழியாகப் பயிலும் மாணவர்களுக்கும் NLP கருவிகள் மிகவும் உதவியாக உள்ளன.
- **தானியங்கி பிழைத்திருத்தம் (Auto-correction):** தமிழில் எழுதும் போது ஏற்படும் எழுத்துப் பிழைகள், இலக்கணப் பிழைகளைத் தானாகத் திருத்தும் கருவிகள் கல்வியில் மிகவும் பயனுள்ளவை.
- **தமிழ் வாசிப்பு கருவிகள் (Text-to-Speech):** பார்வையற்றோர் மற்றும் படிக்கத் தெரியாதோருக்காக, தமிழ் உரையைக் குரலாக மாற்றும் கருவிகள் தமிழ்க் கல்வியை எளிமையாக்குகின்றன.
- **தன்னியக்க வினாத்தாள் உருவாக்கம்:** ஒரு பாடத்தைப் பகுப்பாய்வு செய்து, அதிலிருந்து தன்னியக்கமாக வினாத்தாள்கள் மற்றும் பயிற்சி கேள்விகளை உருவாக்கும் NLP கருவிகள் ஆசிரியர்களுக்கு நேரத்தைச் சேமிக்கின்றன.

முடிவுரை

இயற்கை மொழி செயலாக்கம் (NLP) தமிழ் மொழியின் எதிர்காலத்தை மறுவடிவமைக்கும் ஒரு சக்திவாய்ந்த தொழில்நுட்பமாக விளங்குகிறது. இது இயந்திர மொழிபெயர்ப்பு, பேச்சு அங்கீகாரம், உரைச் சுருக்கம், உணர்வு பகுப்பாய்வு, தேடுபொறிகள், மற்றும் கல்விக் கருவிகள் எனப் பல்துறைகளில் பயனுள்ள பங்களிப்புகளை வழங்குகிறது. இருப்பினும், தரவுப் பற்றாக்குறை, ஒட்டுக்கொள்ளும் மொழியியல், இருமொழி நிலை, மற்றும் தரப்படுத்தப்பட்ட குறியீட்டு முறை இல்லாமை ஆகியவை பெரும் சவால்களாகத் தொடர்ந்து நிலவுகின்றன. இந்தச் சவால்களைச் சமாளிப்பதற்கு, கல்வி நிறுவனங்கள், தொழில்நுட்ப நிறுவனங்கள், மற்றும் மொழியியல் ஆய்வாளர்கள் ஒன்றிணைந்து, பெரிய தரவுத்

தொகுப்புகளை உருவாக்குதல், புதிய ஆழ்ந்த கற்றல் மாதிரிகளை வடிவமைத்தல், மற்றும் நெறிமுறைக் கட்டமைப்புகளை உருவாக்குதல் ஆகியவை இன்றியமையாதவை. தமிழ் மொழியின் தொன்மையான அழகியலையும், நவீன தொழில்நுட்பத்தின் ஆற்றலையும் இணைக்கும் இந்த முயற்சி, தமிழ் மொழியை உலகளாவிய டிஜிட்டல் மொழியாக மாற்றும் ஒரு பெரிய பாய்ச்சலாக அமையும். NLP-யின் முன்னேற்றத்திற்கும், தமிழ் மொழியின் பாதுகாப்பு மற்றும் வளர்ச்சிக்கும் இடையே உள்ள இந்த இணைவு, தமிழ்க் கல்வி, ஆய்வு, மற்றும் எழுத்துத் துறைகளில் ஒரு புதிய சகாப்தத்தைத் திறக்கும் என்பதில் ஐயமில்லை.

குறிப்புகள்

- [1] சிவப்பிரகாசம், க. (2021). தமிழ் மொழியியலும் கணினி மொழியியலும்: ஒரு அறிமுகம். சென்னை: தமிழ்ப் பல்கலைக்கழக வெளியீட்டகம்.
- [2] குமரேசன், பி. (2022). இயற்கை மொழி செயலாக்கத்தின் அடிப்படைகள் மற்றும் தமிழ்ப் பயன்பாடுகள். கோவை: கணினி மொழி வெளியீடு.
- [3] பாலகிருஷ்ணன், ரா. (2022). தமிழ் உணர்வு பகுப்பாய்வில் சொற்பொருள் சவால்கள். இந்திய மொழி மற்றும் தொழில்நுட்ப ஆய்வுகள், 10(4), 45-62.
- [4] மேன்னிங், சி. டி., & ஷூட்சே, எச். (2021). புள்ளிவிவர இயற்கை மொழி செயலாக்கத்தின் அடித்தளங்கள். மொழிபெயர்ப்பு: சண்முகம், கே. சென்னை: நியூ செஞ்சுரி புத்தக நிலையம்).
- [5] ஜுராஃப்ஸ்கி, டி., & மார்ட்டின், ஜே. எச். (2023). பேச்சு மற்றும் மொழிச் செயலாக்கம். மொழிபெயர்ப்பு: குமரேசன், பி. கோவை: கணினி மொழி வெளியீடு).
- [6] தமிழ் NLP சமூகம். (2025). தமிழ் தரவுத் தொகுப்புகள் மற்றும் கருவிகளின் களஞ்சியம். (இணைய ஆவணம்). கிடைக்கும் இடம்: www.tamilnlp.org/datasets. அணுகல் நாள்: 02 ஆகஸ்ட் 2026.

நிதிசார் கட்டுரையாளர் உறுதிமொழி: இல்லை

கட்டுரையாளர் நன்றியுரை: இல்லை

கட்டுரையாளர் உறுதிமொழி: இக்கட்டுரையில் எவ்வித முரண்பாடும் இல்லை என்று உறுதிமொழி அளிக்கிறேன்.



இக்கட்டுரை கிரியேட்டிவ் காமன்சு ஆட்ரிபியூசன் 4.0வின் [Creative Commons Attribution4.0](https://creativecommons.org/licenses/by/4.0/) கீழ் பன்னாட்டு உரிமம் பெற்றுள்ளது.