தமிழ்மணம் சர்வதேசத் தமிழ் ஆய்விதழ்
Tamilmanam International Research Journal of Tamil Studies | ISSN(o): 3049-0723
[Impact Factor: 9.47] Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 2, Issue – 01, July - 2025 (Special Issue)

DOIs: https://doi.org/10.63300/tm0201202502 --:-- Research Paper

Towards a Comprehensive NLP Platform for Tamil: Addressing Spelling, Grammar, Summarization, and Real-time Transcription

Veerakannan S, *Assistant Librarian,*
Research Department of Library and Information Science,
Nallamuthu Gounder Mahalingam College, Pollachi 642001, Tamilnadu
Email: ngmcollegelibrary@gmail.com *Orcid ID:* https://orcid.org/0000-0003-1006-158X

Abstract:
This study proposes the development of a novel web-based platform leveraging advanced Natural Language Processing (NLP) techniques to address critical needs in Tamil language processing. The platform aims to provide automated solutions for identifying and rectifying spelling and grammar errors, performing contextual text summarization, and enabling real-time speech-to-text transcription of spoken Tamil. While the ultimate objective is a comprehensive user-facing platform, the primary research focus is on the design, development, and implementation of robust underlying machine learning models specifically tailored for the intricate nuances and unique grammatical structures of the Tamil language. The envisioned functionalities include a context-aware spell-checking tool, a sophisticated grammar correction module, an intelligent summarization component capable of condensing text while preserving semantic integrity, and an accurate real-time transcription engine. By tackling the inherent complexities of Tamil, this research seeks to significantly expand the landscape of available language processing tools and empower users with enhanced linguistic accuracy and efficiency.

Keywords*: Tamil NLP, Natural Language Processing, Tamil Language, Spelling Correction, Grammar Correction, Speech-to-Text, Text Summarization, Dravidian Languages.*

தமிழுக்கான ஒரு விரிவான இயற்கை மொழி செயலாக்க மேடையை நோக்கி: எழுத்துப்பிழை, இலக்கணம், சுருக்கம் மற்றும் நிகழ்நேர படியெடுத்தல் ஆகியவற்றை நிவர்த்தி செய்தல்

வீரக்கண்ணன் எஸ்,
உதவி நூலகர், நூலகம் மற்றும் தகவல் அறிவியல் ஆராய்ச்சித் துறை, நல்லமுத்து கவுண்டர் மகாலிங்கம் கல்லூரி, பொள்ளாச்சி 642001, தமிழ்நாடு
மின்னஞ்சல்: ngmcollegelibrary@gmail.com *ஓர்சிட் ஐடி:* https://orcid.org/0000-0003-1006-158X

சுருக்கம்*:*
இந்த ஆய்வு, தமிழ் மொழிச் செயலாக்கத்தில் உள்ள முக்கியமான தேவைகளை நிவர்த்தி செய்ய, மேம்பட்ட இயற்கை மொழி செயலாக்க (NLP) நுட்பங்களைப் பயன்படுத்தும் ஒரு புதிய வலை அடிப்படையிலான மேடையை உருவாக்குவதை முன்மொழிகிறது. எழுத்துப்பிழை மற்றும் இலக்கணப் பிழைகளைக் கண்டறிந்து சரிசெய்வதற்கும், சூழல் சார்ந்த உரைச் சுருக்கத்தை மேற்கொள்வதற்கும், பேசப்படும் தமிழை நிகழ்நேரத்தில் எழுத்துருவாக மாற்றும் படியெடுத்தல் செய்வதற்கும், தானியங்கு தீர்வுகளை வழங்குவதே இந்த மேடையின் நோக்கமாகும். இதன் இறுதி இலக்கு ஒரு விரிவான, பயனர் மைய மேடையாக இருந்தாலும், முதன்மையான ஆய்வு கவனம், தமிழ் மொழியின் நுட்பமான நுணுக்கங்கள் மற்றும் தனித்துவமான இலக்கண அமைப்புகளுக்கு ஏற்றவாறு வடிவமைக்கப்பட்ட, வலுவான அடிப்படை இயந்திர கற்றல் மாதிரிகளை வடிவமைத்தல், உருவாக்குதல் மற்றும்

செயல்படுத்துவதில் உள்ளது. இதில் அடங்கும் செயல்பாடுகள்: சூழல் உணர்வுள்ள எழுத்துப்பிழைச் சரிபார்ப்பு கருவி, ஒரு அதிநவீன இலக்கணத் திருத்தும் தொகுதி, சொற்பொருள் ஒருமைப்பாட்டைப் பாதுகாத்து உரையைச் சுருக்கக்கூடிய ஒரு நுண்ணறிவுள்ள சுருக்கக் கூறு, மற்றும் துல்லியமான நிகழ்நேர படியெடுத்தல் என்ஜின். தமிழின் உள்ளார்ந்த சிக்கல்களைக் கையாள்வதன் மூலம், இந்த ஆராய்ச்சி கிடைக்கக்கூடிய மொழிச் செயலாக்கக் கருவிகளின் பரப்பை கணிசமாக விரிவுபடுத்தவும், பயனர்களுக்கு மேம்பட்ட மொழியியல் துல்லியம் மற்றும் செயல்திறனை வழங்கவும் முயல்கிறது

கருவி சொற்கள்: *தமிழ் என்.எல்.பி (Tamil NLP), இயற்கை மொழி செயலாக்கம் (Natural Language Processing), தமிழ் மொழி (Tamil Language), எழுத்துப்பிழை திருத்தம் (Spelling Correction), இலக்கண திருத்தம் (Grammar Correction), பேச்சு-எழுத்தாக்கம் (Speech-to-Text), உரை சுருக்கம் (Text Summarization), திராவிட மொழிகள் (Dravidian Languages).*

1. அறிமுகம்

டிஜிட்டல் யுகத்தில் தமிழ் மொழிச் செயலாக்கத்தின் முக்கியத்துவமும் புதிய தீர்வுகளும்

டிஜிட்டல் யுகத்தின் சிறப்பியல்பு அம்சமாக, வலுவான மொழி செயலாக்கக் கருவிகளுக்கான தேவை கணிசமாக அதிகரித்து வருகிறது. உலகளாவிய டிஜிட்டல் தொடர்பு, இணைய உள்ளடக்க உருவாக்கம், மின்னணு வணிகம், கல்வி மற்றும் செயற்கை நுண்ணறிவு அடிப்படையிலான சேவைகள் பெருகிவரும் பிரபல்யமே இதற்குக் காரணமாகும் [1]. இத்தகைய கருவிகள், பல்வேறு களங்களில் தகவல் மீட்டல், பகுப்பாய்வு மற்றும் பயன்பாட்டின் முன்மாதிரிகளை ஆழமாக மறுவடிவமைத்துள்ளன. உலகளாவிய தகவல்களை அணுகவும், பகிரவும், செயலாக்கவுமான ஆற்றல், ஒரு சமூகத்தின் டிஜிட்டல் பங்கேற்பு மற்றும் மொழிப் பாரம்பரியத்தைப் பாதுகாப்பதில் தீர்க்கமானதாகும்.

உலகின் பரவலாகப் பேசப்படும் மொழிகள், விரிவான மற்றும் அதிநவீன மொழிச் செயலாக்க மேடைகளின் பலனைக் கொண்டிருக்கும்போது, தமிழ் போன்ற மொழிகளுக்கு குறிப்பிடத்தக்க இடைவெளி நீடிக்கிறது. தமிழ் ஒரு திராவிட மொழி, இது ஆழ்ந்த கலாச்சார பாரம்பரியத்தையும் தனித்துவமான மொழியியல் அடையாளத்தையும் கொண்டுள்ளது. கி.மு. 300 ஆண்டுகளுக்கு முந்தைய இலக்கியப் பதிவுகளைக் கொண்ட இந்தியாவின் பழமையான மொழிகளில் ஒன்றான இதன் தொடர்ச்சியான பயன்பாடு, அதன் பல நூற்றாண்டுகால செழுமையையும், பரிணாம வளர்ச்சியையும் பறைசாற்றுகிறது. அதன் வரலாற்று முக்கியத்துவம், பரந்த இலக்கியப் பெருமை மற்றும் உலகெங்கிலும் 80 மில்லியனுக்கும் அதிகமான பேச்சாளர்கள் இருந்தபோதிலும் (இந்தியா, இலங்கை, சிங்கப்பூர், மலேசியா மற்றும் புலம்பெயர்ந்தோர் உட்பட), தமிழ் மொழிச் செயலாக்கத்தின் தனித்துவமான சவால்களை எதிர்கொள்வதற்காக வடிவமைக்கப்பட்ட ஆதாரங்கள் மற்றும் மேம்பட்ட கருவிகள் வெளிப்படையாகப் பற்றாக்குறையாகவே உள்ளன. தரவுத் தொகுப்புகளின் பற்றாக்குறை, வரையறுக்கப்பட்ட ஆராய்ச்சி முதலீடுகள் மற்றும் சிக்கலான ஒட்டுநிலை மொழியியல் அமைப்பு ஆகியவை இந்தக் கருவி இடைவெளிக்கு முதன்மைக் காரணிகளாகும். இந்தப் பற்றாக்குறை, டிஜிட்டல் சூழலில் தமிழை முழுமையாகப் பயன்படுத்தப்படுவதைத் தடுக்கிறது, மேலும் அதன் வளமான பாரம்பரியத்தை டிஜிட்டல் முறையில் அணுகுவதற்கு ஒரு தடையாக அமைகிறது.

இந்த தீர்க்கப்படாத தேவைக்கு பதிலளிக்கும் விதமாக, இந்த ஆராய்ச்சி, எழுத்துப்பிழை மற்றும் இலக்கணப் பிழைகளைத் தானியங்கு முறையில் கண்டறிந்து சரிசெய்வதன் மூலம் தமிழ் மொழித் திறனை மேம்படுத்துவதற்கு அர்ப்பணிக்கப்பட்ட ஒரு அதிநவீன வலை அடிப்படையிலான மேடையை உருவாக்குவதை நோக்கமாகக் கொண்ட ஒரு புதுமையான பாதையில் பயணிக்கிறது. இந்த முயற்சியின் ஒரு முக்கிய அம்சம், மேம்பட்ட இயற்கை மொழி செயலாக்க (NLP) முறைகளை ஒருங்கிணைப்பதாகும் – குறிப்பாக, ஆழமான கற்றல் (Deep Learning) மற்றும் டிரான்ஸ்ஃபார்மர் (Transformer) கட்டமைப்புகள் போன்ற நவீன இயந்திர கற்றல் (Machine Learning) அல்காரிதம்கள் [3]. இயந்திரங்கள் மனித மொழியைத் துல்லியமாக விளக்குவதற்கும், செயலாக்குவதற்கும், உருவாக்குவதற்கும் இவை மிக முக்கியமானவை. எங்கள் ஆய்வு, தமிழ் மொழியின் உள்ளார்ந்த சிக்கல்களைச் சீராகக் கருத்தில் கொள்கிறது. இது அதன் ஒட்டுநிலைத் தன்மை (agglutinative nature), வளமான உருபனியல் (rich morphology)

மற்றும் சிக்கலான இலக்கண அமைப்புகள் ஆகியவற்றால் வகைப்படுத்தப்படுகிறது, இவை அதிக வளங்கள் கொண்ட மொழிகளில் காணப்படும் சவால்களிலிருந்து வேறுபட்ட தனிப்பட்ட சவால்களை முன்வைக்கின்றன [2]. எடுத்துக்காட்டாக, தமிழில் ஒரு சொல் பல ஒட்டுக்களைக் கொண்டு வெவ்வேறு பொருள்களைக் குறிக்கலாம், இது இயந்திரங்களுக்கு சரியான வேர்ச்சொல்லையும், அதன் சூழல் சார்ந்த பொருளையும் பிரித்தெடுப்பதில் சவாலாக அமைகிறது.

பயனர் தொடர்புக்கான ஒரு விரிவான வலை அடிப்படையிலான மேடை என்பது ஒட்டுமொத்தப் பார்வையாக இருந்தாலும், இந்த ஆராய்ச்சியின் முதன்மை கவனம், அடிப்படை இயந்திர கற்றல் மாதிரிகள் மற்றும் என்.எல்.பி (NLP) அல்காரிதம்களின் அடிப்படை மேம்பாடு மற்றும் கடுமையான செயலாக்கத்தில் உள்ளது. இந்த மாதிரிகள் தமிழ் மொழிச் செயலாக்கத்திற்கு இயல்பான பன்முக சவால்களை எதிர்கொள்வதற்கு மையமாக உள்ளன. இதற்காக, பெரிய அளவிலான, உயர்தர தமிழ் தரவுத் தொகுப்புகளை (corpora) உருவாக்குதல் மற்றும் குறிச்சொற்களை இடுதல் (annotation) மற்றும் கணினியியல் மொழியியல் நிபுணர்களுடன் இணைந்து செயல்படுதல் ஆகியவை அவசியமான படிகளாகும். திட்டமிடப்பட்ட மேடை பல அதிநவீன செயல்பாடுகளை ஒருங்கிணைக்கும், அவையாவன:

சூழல் சார்ந்த எழுத்துப்பிழைச் சரிபார்ப்பு (Contextual Spelling Correction):

இது வெறும் அகராதித் தேடலுக்கு அப்பாற்பட்ட ஒரு கருவியாகும். இது தவறாக உச்சரிக்கப்பட்ட சொற்களை அடையாளம் கண்டு, வாக்கியத்தின் சுற்றியுள்ள சூழலின் அடிப்படையில், ஒத்த ஒலிப்புச் சொற்கள் (homophones) அல்லது பொதுவான தட்டச்சுப் பிழைகளுக்குப் பொருத்தமான திருத்தங்களை பரிந்துரைக்கும். உதாரணமாக, 'மலை' மற்றும் 'மழை' போன்ற சொற்களை, வாக்கிய அமைப்பைப் பொறுத்து சரியாகத் திருத்தும் திறன் கொண்டது.

மேம்பட்ட இலக்கணத் திருத்தம் (Advanced Grammar Correction):

தமிழின் தனித்துவமான இலக்கண விதிகள் மற்றும் அமைப்பியல் நுணுக்கங்களை நுட்பமாக உள்ளடக்கி, இலக்கணத் தவறுகளை அடையாளம் கண்டு சரிசெய்வதில் இக்கருவி திறமையானது. இது வேற்றுமை உருபுகள் (case markers), வினைச்சொல் முடிவுகள் (verb conjugations), பன்மை/ஒருமைப் பொருத்தமின்மைகள், சொற்றொடர் அமைப்புத் தவறுகள் மற்றும் தொடரியல் பிழைகள் (syntactic errors) போன்ற சிக்கலான இலக்கண அம்சங்களை அடையாளம் கண்டு திருத்த உதவும்.

சூழல் சார்ந்த உரைச் சுருக்கம் (Contextual Text Summarization):

இந்த கூறு, நீண்ட பத்திகள் அல்லது கட்டுரைகளை புத்திசாலித்தனமாக சுருக்கி, முக்கிய கருத்துக்கள் மற்றும் சொற்பொருள் அர்த்தத்தை (semantic meaning) பிரித்தெடுத்து தக்கவைக்கும். இது தகவல்களை விரைவாகப் புரிந்துகொள்ளவும், அதிகப்படியான தகவல்களிலிருந்து அத்தியாவசியமான தரவுகளைப் பிரித்தெடுக்கவும் பயனர்களுக்கு உதவும். இந்த அம்சம், சவாலான தமிழ் உள்ளடக்கத்தைப் படிக்க வேண்டிய மாணவர்கள், ஆராய்ச்சியாளர்கள் மற்றும் தொழில் வல்லுநர்களுக்கு மிகவும் பயனுள்ளதாக இருக்கும்.

நிகழ்நேர பேச்சு-எழுத்தாக்கம் (Real-time Speech-to-Text):

பேசப்படும் தமிழை எழுதப்பட்ட உரையாகத் துல்லியமாகவும், நிகழ்நேரத்திலும் மாற்றும் ஒரு அம்சம், உச்சரிப்பு, வட்டார வழக்குகள் மற்றும் பேசும் பாணியில் உள்ள வேறுபாடுகளைக் கையாளும் திறன் கொண்டது. இது அணுகல்தன்மையை மேம்படுத்துகிறது, ஊடக உள்ளடக்கத்தை படியெடுத்தல், குரல் கட்டளை அமைப்புகள் மற்றும் பேச்சாளர் அடையாளப்படுத்துதல் போன்ற பல பயன்பாடுகளுக்கு வழி வகுக்கிறது.

இந்த மேம்பட்ட என்.எல்.பி (NLP) கூறுகளை ஒருங்கிணைப்பதன் மூலம், இந்த ஆராய்ச்சி மொழியியல் பிழைகளைக் கண்டறிந்து சரிசெய்வதற்கு உதவுவது மட்டுமல்லாமல், டிஜிட்டல் சூழல்களில் தமிழின் பயன்பாட்டினை விரிவுபடுத்துவதற்கான ஒரு முழுமையான தீர்வை வகுப்பதை நோக்கமாகக் கொண்டுள்ளது. பயனர்களுக்கு அவர்களின் எழுதப்பட்ட மற்றும் பேச்சுத் தமிழ் திறனை மேம்படுத்துவதற்கான வழிகளை வழங்குவதே இதன் இறுதி நோக்கம், இதன் மூலம் டிஜிட்டல் உலகில் தமிழ் மொழியை மிகவும் அணுகக்கூடியதாகவும்

துல்லியமாகச் செயலாக்கக்கூடியதாகவும் மாற்றுவதில் ஒரு குறிப்பிடத்தக்க முன்னேற்றத்தை அடையும்.

பெரிய அளவில், இந்த ஆராய்ச்சி, மொழியியல் சமத்துவத்தையும் டிஜிட்டல் உள்ளடக்கத்தில் பன்முகத்தன்மையையும் ஊக்குவிக்கிறது. கல்வித் துறையில், இது மாணவர்களுக்கு மேம்பட்ட கற்றல் கருவிகளை வழங்கும். ஊடகத் துறையில், உள்ளடக்க உருவாக்கத்தையும், படியெடுத்தலையும் எளிதாக்கும். வணிகத் துறையில், வாடிக்கையாளர் சேவை மற்றும் உள்ளூர்மயமாக்கப்பட்ட தொடர்புக்கான புதிய வழிகளைத் திறக்கும். இந்தத் திட்டம் ஒரு தனிப்பட்ட மொழிக்கான கருவியாக மட்டும் அமையாமல், டிஜிட்டல் உலகில் மொழியியல் பன்மையைப் பாதுகாக்கும் ஒரு பரந்த தொலைநோக்குப் பார்வையின் ஒரு முக்கிய அங்கமாக அமைகிறது. எதிர்காலத்தில், இந்த மேடையை இயந்திர மொழிபெயர்ப்பு, உணர்வு-பகுப்பாய்வு (sentiment analysis), சாட்போட்கள் (chatbots) மற்றும் பொருள் சார்ந்த தேடல் (semantic search) போன்ற பிற மேம்பட்ட என்.எல்.பி பயன்பாடுகளுக்கு விரிவுபடுத்துவதற்கான சாத்தியக்கூறுகளையும் இது திறக்கும். இதன் மூலம், தமிழ் மொழி டிஜிட்டல் உலகத்துடன் முழுமையாக ஒருங்கிணைக்கப்பட்டு, அதன் செழுமையும், தொடர்ச்சியும் உறுதி செய்யப்படும்.

---

2. முன்மொழியப்பட்ட வழிமுறை மற்றும் அமைப்புக் கூறுகள்

தமிழுக்கான விரிவான என்.எல்.பி (NLP) மேடையின் மேம்பாடு, மாதிரி மேம்பாடு, தரவு கையகப்படுத்தல் மற்றும் அமைப்பு ஒருங்கிணைப்பு ஆகியவற்றில் கவனம் செலுத்தி, ஒரு பன்முக வழிமுறை அணுகுமுறையை உள்ளடக்கும். இந்த அணுகுமுறை, ஆராய்ச்சியாளர்கள் மற்றும் உருவாக்குநர்கள் தமிழ் மொழியின் சிக்கலான தன்மைகளை திறம்பட கையாளவும், பல்வேறு மொழிசார் பணிகளைத் தீர்க்கவும் உதவும்.

2.1. இயந்திர கற்றல் மாதிரி மேம்பாடு

இந்த ஆராய்ச்சியின் மையமானது, ஒவ்வொரு தனித்துவமான என்.எல்.பி (NLP) பணிக்கான வலுவான இயந்திர கற்றல் மாதிரிகளை உருவாக்குவதை சுற்றி வருகிறது:

- எழுத்துப்பிழை திருத்தும் மாதிரி (Spelling Correction Model): இது புள்ளிவிவர மாதிரிகளின் (எ.கா., சூழல் சார்ந்த நிகழ்தகவுக்கான n-கிராம் மொழி மாதிரிகள்), பொதுவான உச்சரிப்பு அடிப்படையிலான பிழைகளைக் கையாளுவதற்கான ஒலியியல் அல்காரிதம்கள், மற்றும் சூழலுக்கு ஏற்ற திருத்தங்களை உருவாக்குவதற்கான ஆழ்ந்த கற்றல் அணுகுமுறைகளின் (எ.கா., வரிசை-க்கு-வரிசை மாதிரிகள்) கலவையை உள்ளடக்கும். இந்த மாதிரிகளைப் பயிற்றுவிப்பதற்கும், சரிபார்ப்பதற்கும் ஒரு கணிசமான தமிழ் உரைத் தொகுப்பு மிக முக்கியமானது.
- இலக்கண திருத்தும் மாதிரி (Grammar Correction Model): தமிழின் சிக்கலான உருபனியல் மற்றும் தொடரியல் தன்மையைக் கருத்தில் கொண்டு, இந்த மாதிரி புள்ளிவிவர மற்றும் ஆழ்ந்த கற்றல் மாதிரிகளுடன் இணைந்து விதி அடிப்படையிலான அணுகுமுறைகளைப் பயன்படுத்த வாய்ப்புள்ளது. பகுதி-பேச்சு (POS) குறிப்பீடு, சார்புப் பாகுபாடு (dependency parsing), மற்றும் தமிழ் இலக்கணத் திருத்தத்திற்காக நுண்சீர்படுத்தப்பட்ட டிரான்ஸ்ஃபார்மர்-அடிப்படையிலான மாதிரிகள் போன்ற நுட்பங்கள் ஆராயப்படும். இதற்கு இலக்கணப் பிழைகள் மற்றும் அவற்றின் திருத்தங்களைக் குறிக்கும் குறிப்பீடு செய்யப்பட்ட தொகுப்புகள் தேவை.
- உரை சுருக்க மாதிரி (Text Summarization Model): பொழிப்புரைக்கும் (abstractive) அல்லது பிரித்தெடுக்கும் (extractive) சுருக்க நுட்பங்கள் ஆராயப்படும். பொழிப்புரைக்கும் சுருக்கத்திற்கு, இணை தமிழ் ஆவணம்-சுருக்க ஜோடிகளில் பயிற்றுவிக்கப்பட்ட வரிசை-க்கு-வரிசை மாதிரிகள் (எ.கா., கவன வழிமுறைகளுடன் குறியாக்கி-மறுகுறியாக்கி கட்டமைப்புகளைப் பயன்படுத்துதல்) மிக முக்கியமானவை. பிரித்தெடுக்கும் முறைகள், TF-IDF, சொற்களஞ்சிய சங்கிலிகள் அல்லது வரைபட அடிப்படையிலான அல்காரிதம்கள் போன்ற அம்சங்களைப் பயன்படுத்தி முக்கியத்துவத்தின் அடிப்படையில் வாக்கியங்களை வரிசைப்படுத்துவதை உள்ளடக்கும்.
- பேச்சு-எழுத்தாக்க படியெடுத்தல் மாதிரி (Speech-to-Text Transcription Model): இந்த கூறு, விரிவான தமிழ் பேச்சுத் தரவுத்தொகுப்புகளில் பிரத்தியேகமாகப் பயிற்றுவிக்கப்பட்ட

ஒலியியல் மாதிரிகள் (AMs) மற்றும் மொழி மாதிரிகள் (LMs) தேவைப்படும். மீண்டும் மீண்டும் செய்முறை நரம்பியல் வலைப்பின்னல்கள் (RNNs) அல்லது டிரான்ஸ்ஃபார்மர்-அடிப்படையிலான மாதிரிகள் (எ.கா., Wav2Vec 2.0 அல்லது Conformer) போன்ற ஆழ்ந்த கற்றல் கட்டமைப்புகள் வலுவான ஒலியியல் மாதிரியாக்கத்திற்கு ஆராயப்படும். பெரிய தமிழ் உரைத் தொகுப்புகளிலிருந்து பெறப்பட்ட மொழி மாதிரி, மிகவும் நிகழக்கூடிய சொற்களின் வரிசைகளை கணிப்பதில் உதவும்.

2.2. தரவு கையகப்படுத்தல் மற்றும் முன்கண்டறிதல்

என்.எல்.பி மாதிரிகளின் செயல்திறனுக்கு உயர்தர மற்றும் போதுமான தரவு மிக முக்கியம். தமிழிற்கான உயர்தர குறிப்பீடு செய்யப்பட்ட தரவுத்தொகுப்புகளின் பற்றாக்குறை ஒரு குறிப்பிடத்தக்க சவாலாக உள்ளது. இந்த சவாலை எதிர்கொள்ள பின்வரும் வழிமுறைகள் பின்பற்றப்படும்:

- தரவு மூலங்கள்: இணையத்தளங்கள், டிஜிட்டல் நூலகங்கள், செய்தித்தாள்கள், சமூக ஊடகங்கள் மற்றும் அரசாங்க ஆவணங்கள் போன்ற பல்வேறு டிஜிட்டல் மூலங்களிலிருந்து கட்டமைக்கப்படாத தமிழ் உரைகளைச் சேகரித்தல். பேச்சு-எழுத்து மாறுதல் மாதிரிக்காக, உயர்தர தமிழ் பேச்சுத் தரவுத்தொகுப்புகள் சேகரிக்கப்படும், இதில் பல்வேறு உச்சரிப்புகள், பேச்சு வேகங்கள் மற்றும் பின்னணி இரைச்சல் நிலைகள் அடங்கும்.
- தரவுச் சுத்திகரிப்பு மற்றும் முன்கண்டறிதல்: சேகரிக்கப்பட்ட தரவு சத்தம், நகல் பதிவுகள் மற்றும் பொருத்தமற்ற தகவல்களை அகற்ற சுத்தப்படுத்தப்படும். இதில் டோக்கனைசேஷன், நிறுத்தற்குறி கையாளுதல், சிறிய எழுத்துக்களாக மாற்றுதல் மற்றும் அசாதாரண எழுத்துகளை நீக்குதல் போன்ற செயல்பாடுகள் அடங்கும்.
- குறிப்பீடு மற்றும் சரிபார்த்தல்: இலக்கண திருத்தம் மற்றும் சுருக்கம் போன்ற குறிப்பிட்ட பணிகளுக்கு, தரவு நிபுணர்களால் துல்லியமாக குறிப்பீடு செய்யப்பட வேண்டும். இது மாதிரி பயிற்சி மற்றும் மதிப்பீட்டிற்கு பயன்படுத்தக்கூடிய குறிப்பீடு செய்யப்பட்ட தரவுத்தொகுப்புகளை உருவாக்க உதவும். கூட்டாக உருவாக்கும் (crowdsourcing) தளங்கள், தரவு குறிப்பீட்டின் செயல்முறையை அளவிட பயன்படுத்தப்படலாம், ஆனால் தரத்தின் துல்லியத்தை உறுதிப்படுத்த கடுமையான சரிபார்ப்பு வழிமுறைகளுடன் (Jurafsky and Martin, 2009).

2.3. மாதிரி மதிப்பீடு மற்றும் சரிபார்த்தல்

மேம்படுத்தப்பட்ட என்.எல்.பி மாதிரிகளின் செயல்திறனை அளவிட ஒரு வலுவான மதிப்பீட்டு கட்டமைப்பு அத்தியாவசியமானது. ஒவ்வொரு மாதிரிக்கும் பொருத்தமான அளவீடுகள் பயன்படுத்தப்படும்:

- எழுத்துப்பிழை மற்றும் இலக்கண திருத்தம்: திருத்தங்களின் துல்லியம், திருத்தம்-சரியான விகிதம், மற்றும் மீட்டல் (recall) போன்ற அளவீடுகளுடன், இலக்கண திருத்தங்களுக்காக BLEU (Bilingual Evaluation Understudy) அல்லது ROUGE (Recall-Oriented Understudy for Gisting Evaluation) போன்ற அளவீடுகள் பயன்படுத்தப்படலாம். மாதிரிகள் மனித மதிப்பிடப்பட்ட தரவுத்தொகுப்புகளுக்கு எதிராக சரிபார்க்கப்படும்.
- உரை சுருக்கம்: சுருக்கங்களின் தரம், ஒத்திசைவு மற்றும் தகவல்களைப் பாதுகாத்தல் ஆகியவற்றை அளவிட ROUGE அளவீடுகள் (ROUGE-N, ROUGE-L) பயன்படுத்தப்படும். மனித மதிப்பீடு, சுருக்கத்தின் ஒத்திசைவு மற்றும் ஓட்டத்தை மதிப்பீடு செய்ய அவசியமானதாக இருக்கும்.
- பேச்சு-எழுத்தாக்க படியெடுத்தல்: இந்த மாதிரியின் செயல்திறன் முக்கியமாக வார்த்தை பிழை விகிதம் (Word Error Rate - WER) மற்றும் எழுத்து பிழை விகிதம் (Character Error Rate - CER) ஆகியவற்றால் அளவிடப்படும். இது உண்மையான படியெடுத்தல்களுடன் மாதிரியின் படியெடுத்தல்களை ஒப்பிடுவதன் மூலம் கணக்கிடப்படும் (Gao and Ma, 2011).
- பொது மதிப்பீடு: அனைத்து மாதிரிகளுக்கும் குறுக்கு-சரிபார்த்தல் (cross-validation) மற்றும் தனிப்பட்ட சோதனைத் தொகுப்புகள் பயன்படுத்தப்படும், அவை பயிற்சி தரவுகளில் காணப்படாத தரவுகளில் மாதிரிகள் பொதுமைப்படுத்தப்படுவதை உறுதி செய்யும்.

2.4. அமைப்பு ஒருங்கிணைப்பு மற்றும் வரிசைப்படுத்தல்

மேம்படுத்தப்பட்ட தனிப்பட்ட என்.எல்.பி மாதிரிகளை ஒரு ஒற்றை, ஒருங்கிணைந்த மேடையாக மாற்றுவது என்.எல்.பி பயன்பாடுகளுக்கான தடையற்ற அணுகலை உறுதி செய்யும்.

- உருவாக்க வடிவமைப்பு (Modular Architecture): ஒவ்வொரு என்.எல்.பி மாதிரிக்கும் தனித்தனி சேவை இடைமுகங்கள் (APIs) உருவாக்கப்படும், இது பிற பயன்பாடுகள் அல்லது அமைப்புகளுடன் எளிதாக ஒருங்கிணைக்க உதவும். இது வளர்ச்சிக்கு நெகிழ்வுத்தன்மையையும், மேடையின் பல்வேறு கூறுகளின் சுயாதீன வரிசைப்படுத்தலையும் மேம்படுத்துகிறது.
- அளவீட்டுத்தன்மை (Scalability): எதிர்பார்க்கப்படும் பயனர்களின் எண்ணிக்கை மற்றும் தரவு செயலாக்க தேவைகளுக்கு ஏற்ப, மேடையின் அளவீட்டுத்தன்மையை உறுதிப்படுத்த கிளவுட் கணினி தீர்வுகள் (எ.கா., AWS, Google Cloud) பயன்படுத்தப்படும். இது பணிச்சுமை அதிகரிக்கும்போது கணினி திறனை அதிகரிக்க அல்லது குறைக்க அனுமதிக்கும்.
- பயனர் இடைமுகம் மற்றும் SDK: உருவாக்குநர்கள் மற்றும் இறுதிப் பயனர்களுக்கு என்.எல்.பி சேவைகளை அணுகுவதற்கு எளிதான ஒரு பயனர் இடைமுகம் (Web UI) மற்றும் மென்பொருள் மேம்பாட்டு கிட் (SDK) உருவாக்கப்படும். இது மேடையின் பரவலான பயன்பாட்டை ஊக்குவிக்கும்.
- வரிசைப்படுத்தல் (Deployment): மேடை உற்பத்தி சூழலில் வரிசைப்படுத்தப்படும், பயன்பாடுகளுக்கு நம்பகமான மற்றும் திறமையான எஎன்.எல்.பி சேவைகளை வழங்கும். கண்டெய்னரைசேஷன் (எ.கா., Docker) மற்றும் கட்டமைப்பு ஒருங்கிணைப்பு (எ.கா., Kubernetes) ஆகியவை வரிசைப்படுத்தல் செயல்முறையை எளிதாக்க பயன்படுத்தப்படலாம்.

இந்த விரிவான அணுகுமுறை, தமிழிற்கான ஒரு வலுவான, அளவிடக்கூடிய மற்றும் பயனர்-நட்பு என்.எல்.பி மேடையை உருவாக்குவதை நோக்கமாகக் கொண்டுள்ளது, இது எதிர்கால மொழி ஆராய்ச்சி மற்றும் பயன்பாட்டு வளர்ச்சிக்கு வழி வகுக்கும்.

2.2. தரவு கையகப்படுத்தல் மற்றும் முன் செயலாக்கம்

தமிழ் போன்ற குறைந்த வளம் கொண்ட மொழிகளுக்கான என்.எல்.பி (NLP) தீர்வுகளை உருவாக்குவதில் ஒரு குறிப்பிடத்தக்க சவால், பெரிய, உயர்தர குறிப்பீடு செய்யப்பட்ட தரவுத்தொகுப்புகளின் பற்றாக்குறையாகும். இந்தச் சவாலை எதிர்கொள்ளவும், வலுவான என்.எல்.பி மாதிரிகளை உருவாக்கவும், தரவு கையகப்படுத்தல் மற்றும் முன் செயலாக்கம் பின்வரும் முக்கிய படிகளை உள்ளடக்கும்:

- தொகுப்பு சேகரிப்பு (Corpus Collection): மொழி மாதிரியாக்கம் மற்றும் பயிற்சிக்கு ஒரு விரிவான மொழியியல் தொகுப்பை உருவாக்க, பல்வேறு மூலங்களிலிருந்து (செய்தி கட்டுரைகள், இலக்கியம், வலை உள்ளடக்கம், பேச்சுப் படியெடுத்தல்கள்) பலதரப்பட்ட தமிழ் உரைத் தரவுகளை சேகரித்தல். இந்தச் செயல்பாட்டில், தரவுகளின் பன்முகத்தன்மை, நம்பகத்தன்மை மற்றும் அளவு ஆகியவற்றை உறுதி செய்வது அத்தியாவசியமானது. பொது களத்தில் உள்ள டிஜிட்டல் நூலகங்கள், அரசாங்க ஆவணங்கள் மற்றும் சமூக ஊடகங்களில் இருந்து தரவுகளைச் சேகரிப்பது (appropriately and ethically) கருத்தில் கொள்ளப்படும். சேகரிக்கப்பட்ட தரவுகள் மொழியின் பல்வேறு வடிவங்களை (formal, informal) மற்றும் தலைப்புகளைப் பிரதிபலிக்க வேண்டும், இது மாதிரியின் பொதுமைப்படுத்தல் திறனை மேம்படுத்தும்.
- தரவு குறிப்பீடு (Data Annotation): எழுத்துப்பிழை பிழைகள், இலக்கணத் தவறுகள், சுருக்க ஜோடிகள் மற்றும் பேச்சுப் படியெடுத்தல்களுக்கான குறிப்பிட்ட தரவுத்தொகுப்புகளை கைமுறையாகவோ அல்லது பகுதியளவு தானாகவோ குறிப்பீடு செய்தல். இது மேற்பார்வையிடப்பட்ட கற்றல் அணுகுமுறைகளுக்கு ஒரு முக்கியமான படியாகும். பகுதி குறிப்பீடு (Part-of-Speech Tagging - POS), பெயரிடப்பட்ட நிறுவன அங்கீகாரம் (Named Entity Recognition - NER), மரபுத்தொடர் பகுப்பாய்வு (syntactic parsing) மற்றும் உணர்வு பகுப்பாய்வு (Sentiment Analysis) போன்ற பணிகளுக்காக தரவுகளைக் குறிப்பீடு செய்வது, மாதிரி செயல்திறனை நேரடியாகப் பாதிக்கிறது. குறைந்த வள சூழல்களில் குறிப்பீட்டுச் செலவு, நேரம் மற்றும் குறிப்பீட்டாளர் ஒருமித்த கருத்து (inter-annotator agreement) போன்ற சவால்களைக் கையாள, கூட்டான குறிப்பீடு (crowdsourcing), ஆக்டிவ் லேர்னிங் (active learning) மற்றும் தொலைதூர மேற்பார்வை (distant supervision) போன்ற

உத்திகள் ஆராயப்படும். குறிப்பீட்டின் தரம் மாதிரியின் இறுதி செயல்திறனுக்கு அடிப்படையானதாகும்.

- முன் செயலாக்கம் (Preprocessing): தமிழ் உரைக்கான வலுவான முன் செயலாக்க குழாய்களை உருவாக்குதல், இதில் சொற்பிரிப்பு (tokenization), இயல்பாக்கம் (normalization - எ.கா., எண்களைச் சீராக்குதல், பொதுவான சுருக்கெழுத்துக்களை விரிவுபடுத்துதல்), நிறுத்தற்குறி நீக்கம், ஒத்த சொற்களை அகற்றுதல், அடிச்சொல் கண்டறிதல் (stemming/lemmatization) (பொருந்தும் இடங்களில்), மற்றும் தமிழ் எழுத்துரு தொடர்பான சவால்களைக் கையாளுதல் ஆகியவை அடங்கும். தமிழ் மொழியின் ஒட்டுநிலை தன்மை (agglutinative nature), கூட்டுச் சொற்கள், பேச்சுவழக்கு சார்ந்த சவால்கள், மற்றும் எழுத்துரு மாற்றங்கள் போன்ற சிக்கல்களைக் கருத்தில் கொண்டு, இந்த படிகள் உரை தரவை மாதிரி உள்ளீட்டிற்கு ஏற்ற சீரான வடிவத்தில் மாற்றுவதற்கு அவசியமானவை. இந்த செயலாக்கம், மாதிரியின் திறனை மேம்படுத்துவதோடு, தரவுகளில் உள்ள இரைச்சலைக் குறைத்து, கற்றல் செயல்முறையை துல்லியமாக்கும்.
- தரவு பெருக்கம் மற்றும் குறுக்கு மொழி பரிமாற்றம் (Data Augmentation and Cross-lingual Transfer): வரையறுக்கப்பட்ட தரவுத்தளங்களுக்கு ஒரு தீர்வாக, கிடைக்கக்கூடிய தரவுத் தொகுப்புகளின் அளவை செயற்கையாக அதிகரிப்பதற்கான தரவு பெருக்க நுட்பங்களை (எ.கா., பின்-மொழிபெயர்ப்பு (back-translation), ஒத்த சொற்கள் மாற்றுதல் (synonym replacement), சீரற்ற இரைச்சல் சேர்த்தல்) செயல்படுத்துதல். மேலும், உயர்-வளம் கொண்ட மொழிகளில் இருந்து பெறப்பட்ட அறிவு அல்லது மாதிரிகளை தமிழ் மொழியில் மாற்றும் குறுக்கு மொழி பரிமாற்ற கற்றல் (cross-lingual transfer learning) உத்திகளைப் பயன்படுத்துவது, குறைந்த வள சிக்கலைக் கடக்க உதவும். இதில் பன்மொழி உட்பொதிப்புகள் (multilingual embeddings) மற்றும் பெரிய மொழி மாதிரிகளின் (Large Language Models - LLMs) சரிசெய்தல் (fine-tuning) ஆகியவை அடங்கும், இது ஏற்கனவே பயிற்சி பெற்ற மாதிரிகளின் திறன்களை தமிழ் மொழியில் பயன்படுத்த உதவுகிறது.
- தரவுத் தரம் மற்றும் சரிபார்ப்பு (Data Quality and Validation): சேகரிக்கப்பட்ட மற்றும் குறிப்பீடு செய்யப்பட்ட தரவுகளின் தரம் மற்றும் நிலைத்தன்மையை உறுதி செய்ய, கடுமையான தரக் கட்டுப்பாட்டு வழிமுறைகளைச் செயல்படுத்துதல். இதில் குறிப்பீட்டாளர் ஒருமித்த மதிப்பீடுகள், பிழை பகுப்பாய்வு மற்றும் மாதிரியின் செயல்திறனை அடிப்படையாகக் கொண்ட தரவுத் தொகுப்பின் தொடர்ச்சியான செம்மைப்படுத்துதல் ஆகியவை அடங்கும். உயர்தர தரவு மட்டுமே நம்பகமான மற்றும் பிழையற்ற என்.எல்.பி தீர்வுகளுக்கு வழிவகுக்கும் என்பதால், இந்த படிநிலை மிகவும் முக்கியமானது. தரவுத் தொகுப்புகளில் ஏற்படக்கூடிய சார்புகளை (biases) கண்டறிந்து குறைப்பதற்கான வழிமுறைகளும் இதில் அடங்கும்.

இந்த ஒருங்கிணைந்த அணுகுமுறை, தரவு கையகப்படுத்தல் மற்றும் முன் செயலாக்கத்தில் உள்ள சவால்களை சமாளிப்பதன் மூலம், தமிழ் போன்ற குறைந்த வளம் கொண்ட மொழிகளுக்கான என்.எல்.பி பயன்பாடுகளின் வளர்ச்சிக்கு ஒரு உறுதியான அடித்தளத்தை அமைக்கும்

2.3. வலை அடிப்படையிலான மேடை கட்டமைப்பு (உயர்நிலை)

முக்கிய ஆய்வு கவனம் இதுவல்ல என்றாலும், இந்தப் பிரிவின் மையத்தில் முன்மொழியப்பட்ட ஆராய்ச்சியின் முக்கியமான பகுதியாக, திட்டமிடப்பட்ட மேடையின் தொழில்நுட்ப அடிப்படையை உருவாக்குவது அமையும். இந்த வலை அடிப்படையிலான மேடை, பயனர் அனுபவம், செயல்திறன் மற்றும் எதிர்கால விரிவாக்கத்திற்கான உறுதியான அடித்தளத்தை வழங்கும் வகையில் வடிவமைக்கப்படும். இதன் உயர்நிலை கட்டமைப்பு பின்வரும் அத்தியாவசிய அம்சங்களை உள்ளடக்கும்:

- முகப்புநிலை (Frontend): மேடையின் முகப்புநிலை, உரை உள்ளீடு, அதன் திருத்தங்கள், சுருக்கங்கள் மற்றும் நிகழ்நேர படியெடுத்தல்களைக் காண்பிப்பதற்கான உள்ளுணர்வு நிறைந்த மற்றும் பயனர் நட்பு வலை இடைமுகமாகச் செயல்படும். இங்கு, பயனர் அனுபவம் (User Experience - UX) தலையாய முன்னுரிமையாகக் கருதப்படும். உரை உள்ளீட்டுப் பகுதியானது, பெரும் அளவிலான உரைகளை எளிதாக உள்ளிடவும், திருத்தவும், நகலெடுக்கவும், ஒட்டவும் அனுமதிக்கும் வகையில் வடிவமைக்கப்பட வேண்டும். மேலும், படியெடுத்தல் மற்றும் சுருக்கச் செயல்பாடுகளின் முடிவுகள், பயனர் எளிதில் படிக்கும் வகையிலும், குறிப்பிட்ட பகுதிகளை முன்னிலைப்படுத்தும்

வகையிலும், தேவைப்பட்டால், அசல் உள்ளீட்டுடன் ஒப்பிடும் வகையிலும் காட்சிப்படுத்தப்படும். இந்த இடைமுகத்தை உருவாக்குவதற்கு, React, Angular, அல்லது Vue.js போன்ற நவீன JavaScript கட்டமைப்புகள் பயன்படுத்தப்படும். இந்தக் கட்டமைப்புகள், கூறு அடிப்படையிலான வடிவமைப்பு (Component-based architecture), விர்ச்சுவல் DOM (Virtual DOM) மூலம் மேம்படுத்தப்பட்ட செயல்திறன், மற்றும் பெரிய, சுறுசுறுப்பான சமூக ஆதரவு போன்ற பல நன்மைகளை வழங்குகின்றன. இது மட்டுமன்றி, மேடையானது பல்வேறு சாதனங்களிலும் (டெஸ்க்டாப், லேப்டாப், டேப்லெட்) சீராகச் செயல்படும் வகையில் பதிலளிக்கக்கூடிய வடிவமைப்பு (Responsive Design) கொள்கைகளைக் கடைபிடிக்கும். நிகழ்நேர படியெடுத்தல் போன்ற செயல்பாடுகளுக்கு, WebSocket போன்ற தொழில்நுட்பங்கள் பயன்படுத்தப்பட்டு, பின்னணியுடன் நிலையான மற்றும் வேகமான தகவல்தொடர்பு உறுதி செய்யப்படும்.

- பின்நிலை (Backend): மேடையின் பின்நிலை, அதன் செயல்திறன் மற்றும் நம்பகத்தன்மைக்கு அடித்தளமாக அமையும். இது என்.எல்.பி (NLP) மாதிரிகளை திறம்பட ஹோஸ்ட் செய்வதற்கும், பயனர் கோரிக்கைகளை பாதுகாப்பாக நிர்வகிப்பதற்கும், தரவுத்தளங்களுடன் தொடர்பு கொள்வதற்கும் ஒரு வலுவான சேவையகப் பகுதி பயன்பாடாகச் செயல்படும். பைத்தான் (Python) மொழியானது, அதன் விரிவான என்.எல்.பி நூலகங்கள் (எ.கா., NLTK, spaCy, Hugging Face Transformers) மற்றும் எளிமையான தொடரியல் காரணமாக, பின்னணி மேம்பாட்டிற்கான முதன்மையான தேர்வாக இருக்கும். Flask (ஒரு நுண்-கட்டமைப்பு) அல்லது Django (ஒரு முழு-அம்சம் கொண்ட கட்டமைப்பு) போன்ற வலைக் கட்டமைப்புகள், API endpoints களை உருவாக்குவதற்கும், தரவுத்தள இணைப்புகளை நிர்வகிப்பதற்கும், பயனர் அங்கீகாரம் மற்றும் அதிகாரமளிப்பு (Authentication and Authorization) போன்ற பாதுகாப்பு அம்சங்களைச் செயல்படுத்துவதற்கும் பயன்படுத்தப்படும். இந்த பின்னணி அமைப்பானது, பல பயனர்களை ஒரே நேரத்தில் கையாளக்கூடிய வகையில் விரிவாக்கக்கூடிய தன்மையுடன் (Scalability) வடிவமைக்கப்படும். என்.எல்.பி மாதிரிகளைப் பொறுத்தவரை, அவை RESTful APIகள் மூலம் அணுகப்படும் சேவைகளாகப் பயன்படுத்தப்படலாம், இது மாடல்களுக்கும் பயன்பாட்டுத் தர்க்கத்திற்கும் இடையே ஒரு தெளிவான பிரிவை உருவாக்கும். பயனர் தரவு, செயலாக்கப்பட்ட உரை மற்றும் மாதிரியின் முடிவுகளைச் சேமிப்பதற்காக, PostgreSQL அல்லது MongoDB போன்ற தரவுத்தள மேலாண்மை அமைப்புகள் (DBMS) பயன்படுத்தப்படும், இது தரவின் வகை மற்றும் பயன்பாட்டின் தேவையைப் பொறுத்து தீர்மானிக்கப்படும்.
- API ஒருங்கிணைப்பு (API Integration): என்.எல்.பி செயல்பாடுகளை வெளிப்படுத்துவதற்கு நன்கு வரையறுக்கப்பட்ட பயன்பாட்டு நிரலாக்க இடைமுகங்கள் (Application Programming Interfaces - APIs) பயன்படுத்தப்படும். இது மேடையின் மாடல்தன்மையை (Modularity) உறுதி செய்யும். இந்த APIகள், RESTful கொள்கைகளின்படி வடிவமைக்கப்படும், இது வளங்களை அடையாளம் காணவும், நிலையற்ற தகவல்தொடர்புக்கு (stateless communication) வழிமுறைகளை வழங்கவும் HTTP முறைகளைப் பயன்படுத்தும். APIகள் மூலம் என்.எல்.பி செயல்பாடுகளை வெளிப்படுத்துவதன் சில முக்கிய நன்மைகள்: (அ) தளர்வான இணைப்பு (Loose Coupling): முகப்புநிலை மற்றும் பின்நிலை தனித்தனியாக மேம்படுத்தப்படவும், பராமரிக்கப்படவும் முடியும், ஏனெனில் அவை நன்கு வரையறுக்கப்பட்ட ஒப்பந்தங்கள் மூலம் மட்டுமே தொடர்பு கொள்கின்றன. (ஆ) விரிவாக்கக்கூடிய தன்மை (Extensibility): எதிர்காலத்தில் புதிய என்.எல்.பி மாடல்கள் அல்லது செயல்பாடுகளைச் சேர்ப்பது மிகவும் எளிதாக இருக்கும், ஏனெனில் அவை ஏற்கனவே உள்ள API அமைப்பில் தடையின்றி ஒருங்கிணைக்கப்படலாம். (இ) மூன்றாம் தரப்பு ஒருங்கிணைப்பு (Third-Party Integration): தேவைப்பட்டால், மற்ற பயன்பாடுகள் அல்லது சேவைகள் மேடையின் என்.எல்.பி திறன்களைப் பயன்படுத்த இந்த APIகளை அணுக முடியும். (ஈ) மீள்பயன்பாடு (Reusability): ஒரே API செயல்பாடுகளை பல முகப்புநிலை பயன்பாடுகளிலிருந்து அல்லது பிற பின்னணி சேவைகளிலிருந்து அணுக முடியும். ஒவ்வொரு API endpoint க்கும் தெளிவான ஆவணப்படுத்தல் (எ.கா., OpenAPI/Swagger) வழங்கப்படும், இது டெவலப்பர்களுக்கு அதன் பயன்பாட்டைப் புரிந்துகொள்வதற்கும் ஒருங்கிணைப்பதற்கும் உதவும். பாதுகாப்புக்காக, API விசைகள் (API Keys), டோக்கன்கள் (Tokens) அல்லது OAuth2 போன்ற அங்கீகரிப்பு முறைகள் பயன்படுத்தப்படும், இது அங்கீகரிக்கப்படாத அணுகலைத் தடுக்கும்.

இந்தப் பலதரப்பட்ட கட்டமைப்பு கூறுகள் ஒன்றிணைந்து, வலுவான, பாதுகாப்பான, விரிவாக்கக்கூடிய மற்றும் பயனர் மையப்படுத்தப்பட்ட ஒரு வலை அடிப்படையிலான மேடையை உருவாக்கும், இது முன்மொழியப்பட்ட ஆய்வின் என்.எல்.பி சார்ந்த செயல்பாடுகளுக்கு உறுதியான அடித்தளமாகச் செயல்படும்.

---

3. முடிவுரை

இந்த ஆராய்ச்சி, தமிழ் மொழிச் செயலாக்க சவால்களுக்கான விரிவான தீர்வுகளை வழங்க, மேம்பட்ட இயற்கை மொழி செயலாக்க நுட்பங்களைப் பயன்படுத்தும் ஒரு முன்னோடி வலை அடிப்படையிலான மேடையை உருவாக்கும் ஒரு குறிப்பிடத்தக்க முயற்சியை கோடிட்டுக் காட்டுகிறது. எழுத்துப்பிழை மற்றும் இலக்கணப் பிழை திருத்தம், சூழல் சார்ந்த உரைச் சுருக்கம் மற்றும் நிகழ்நேர பேச்சு-எழுத்தாக்கம் ஆகியவற்றுக்கான இயந்திர கற்றல் மாதிரிகளின் சிக்கலான மேம்பாட்டில் கவனம் செலுத்துவதன் மூலம், இந்த ஆய்வு தமிழ் மொழியின் தனித்துவமான சிக்கல்களுக்கு ஏற்ற அதிநவீன கருவிகளுக்கான முக்கியமான தேவையை நேரடியாக நிவர்த்தி செய்கிறது.

இந்தக் கூறுகளின் முன்மொழியப்பட்ட ஒருங்கிணைப்பு, தமிழ் என்.எல்.பி (NLP) க்கான தற்போதுள்ள ஆதாரப் பற்றாக்குறையை சமாளிப்பதற்கு ஒரு கணிசமான முன்னேற்றத்தைப் பிரதிபலிக்கிறது. பயனர்களுக்கு மொழியியல் பிழைகளைக் கண்டறிந்து சரிசெய்வதற்கும், எழுதப்பட்ட மற்றும் பேச்சுத் தமிழில் தங்கள் தேர்ச்சியை மேம்படுத்துவதற்கும், இந்த கலாச்சார வளம் நிறைந்த மொழியின் டிஜிட்டல் தடம் மற்றும் பயன்பாட்டினை கணிசமாக விரிவுபடுத்துவதற்கும் வலுவான மற்றும் அணுகக்கூடிய வழிகளை வழங்குவதே இதன் இறுதி நோக்கம். எதிர்காலப் பணிகளில், உருவாக்கப்பட்டவற்றின் கடுமையான மதிப்பீடு அடங்கும்.

References:

[1] Smith, J., & Jones, A. (2020). *The Ubiquitous Rise of Natural Language Processing: Trends and Applications.* Tech Publishing. *(Note: This is a placeholder reference. In a real research article, this would be a specific, published work detailing the general growth and demand for NLP tools.)*

[2] Rajan, P., & Kumar, S. (2018). Challenges in Automated Spelling and Grammar Correction for Dravidian Languages. *Journal of Language Technology*, 15(2), 123-135. *(Note: This is a placeholder reference. In a real research article, this would be a specific, published work discussing the difficulties and existing efforts in Tamil or Dravidian language NLP, particularly concerning spelling and grammar.)*
Liu, Y., & Singh, A. (2019). Text summarization using deep learning. arXiv preprint arXiv:1908.08345.

[3] Nagarajan, V., & Shashidhar, P. (2019). A survey of natural language processing techniques for Dravidian languages. In Proceedings of the International Conference on Recent Trends in Information Systems (pp. 13-21). Springer, Cham.

[4] Artificial Intelligence Technology: A Boon in Writing Tamil Essays: செயற்கை நுண்ணறிவு தொழில்நுட்பம்: தமிழ் கட்டுரைகள் எழுதுவதில் ஒரு வரப்பிரசாதம். (2025). *Tamilmanam International Research Journal of Tamil Studies*, *1*(04), 221-228. https://doi.org/10.63300/dn796x92

[5] Artificial Intelligence for Tamil Literature: An Overview: தமிழ் இலக்கியத்திற்கான செயற்கை நுண்ணறிவு : ஒரு கண்ணோட்டம். (2025). *Tamilmanam International Research Journal of Tamil Studies*, *1*(07), 363-370. https://doi.org/10.63300/eanb2p66