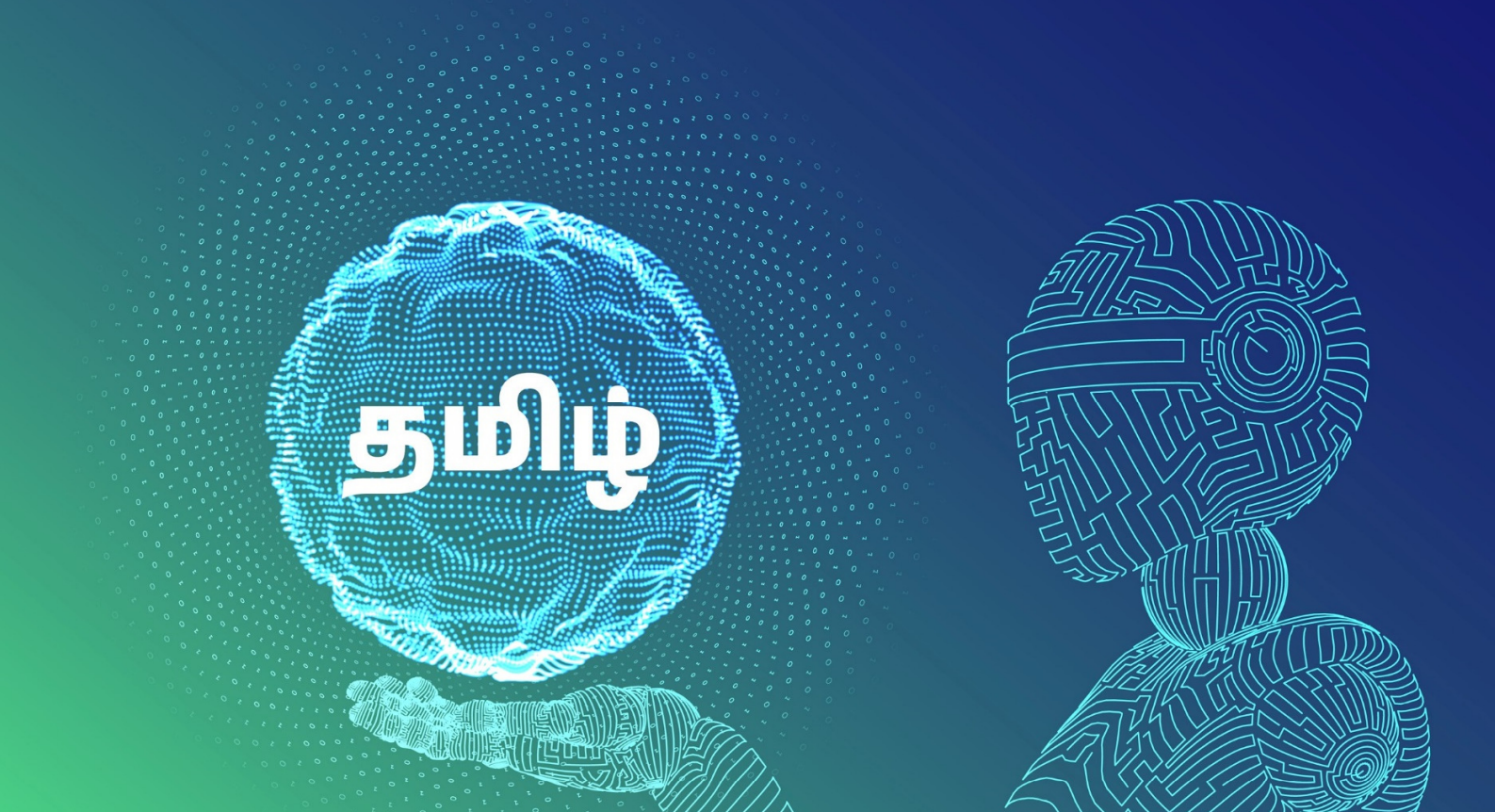
செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தைத் தமிழ்மொழியுடன் செயல்படுத்துவதற்கான பன்னாட்டுக் கருத்தரங்கம்

13 மற்றும் 14 அக்டோபர் 2023
குமரகுரு வளாகம், கோயம்புத்தூர்

மாநாட்டின் கட்டுரைகள்



நா. மகாலிங்கம் தமிழாய்வு மையம்
N.MAHALINGAM TAMIL RESEARCH CENTRE

உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

குமரகுரு

குமரகுரு
நித்திலம்
தேடலின் துவக்கம்

குமரகுரு
கல்வி நிறுவனங்கள்

Greetings from President, Kumaraguru Institutions.

Vanakkam, I am happy to write for the publication that encapsulates the knowledge, insights, and innovative ideas that have flowed through the corridors of this landmark International Conference on Implementation of AI Technologies in Tamil Computing. This book stands as a testament to our collective commitment to advancing the frontiers of technology and preserving the Tamil language. Artificial Intelligence (AI) and the Tamil language have converged in a fascinating journey that combines tradition with cutting-edge technology. This convergence has the potential to reshape, how we interact with and preserve not just the world's oldest language, but also one of the few languages that represents a complete culture with Iyal, Isai and Nadagam, and also with very rare spiritual contents. The conference has been an extraordinary journey, a convergence of some of the brightest minds and dedicated individuals in the fields of AI, Linguistics, and Tamil. It must have ignited the torch of innovation and exploration, redefining the possibilities of Tamil Computing in the AI era.

I am also reminded of the contributions of our visionary founder, Dr. N. Mahalingam to Tamil. Dr. Mahalingam's unwavering passion for the Tamil language and culture has left an indelible mark on the history of the language. The book we release today is a manifesto of our shared vision. It represents the combined wisdom of scholars, researchers, and industry leaders who have committed themselves to bridge the past and the future while propelling it into the digital age. It reflects the dedication and relentless pursuit of excellence demonstrated by INFITT, who work resolutely to promote Tamil Computing in diverse areas bridging the gap between traditional Tamil and modern technology. I would like to express my appreciation to INFITT and the Kumaraguru team who worked on bringing out this book. It is through these collective efforts that we pave the way for a more vibrant and technologically advanced Tamil Nadu. With this book, we take a piece of this conference's spirit and knowledge with us, as we move forward into uncharted territories. May it continue to inspire us, in our ongoing endeavors to nurture the efforts in Tamil Computing and the AI technologies that drive them. I extend my heartfelt thanks to all of you for your presence and contributions to this landmark conference.

Shri. Shankar Vanavarayar,
President, Kumaraguru Institutions.

Greetings from Director, Kumaraguru School of Innovation

I am very happy to be a part of the INFITT Tamil conference at Kumaraguru Institutions. The wealth of the Tamil language is in its literature and Tamil has been more than just a language for the communitywhich sees it as a way of life. Be it integrity, administration, friendship, love, child development, etc Tamil has everything and it makes it easy forus to use them in real life situations. When a language which has evolvedover the years and has been in usage for thousands of years has to meet thetechnology like AI, it says "Vanakkam AI". I feel AI is more an enabler for the young Tamil speaking community to know about the language's rich history and culture. The modern generation which is more used to the handheld devices should get an opportunity to communicate with the Mother Tamil and understand the beauty of the language and realize why we call Tamil as a "Vazhviyal" (life science) rather than just a language. I wish the conference like these bring researchers from across the country and the world to inspire younger generation to make AI learn Tamil and engage with the Tamil speaking people.

I would say Large Language Model when applied to languages like Tamil, will make it a Largest Language Model and there is no other way to respect and recognize a language which has been through ages, still being relevant and make people to wonder on its beauty. The need for creating LLMs for languages like Tamil are manyfold, it enables to preserve the rich knowledge hidden across the texts and would be able to bring them to the needy based on a colloquial slang-based prompt. People of this generation that misses the cultural values may get the benefit of this modelto help guide them in appropriate ways in various stages of life. Be it the moral stories for the children which is usually given by the grandparents, an interactive system that can tell the meaningof the words and their pronunciation, supporting creative writing skills, translating the content forthe consumption and better understanding, AI can do wonders. The language which has been strengthened by the creation of Sangam by great kings of the past is to be driven by AI in future.
...

Vaazhiya Senthamil, Vaazhga Bharatham.

We wish the team all success in their future endeavours.

Dr. Raghuveer V R,
Director, Kumaraguru School of Innovation.

Greetings from Principal, KCLAS

It is with immense pleasure that I extend my warmest greetings on behalf of Kumaraguru College of Liberal Arts and Science. We are delighted to host the Conference on AI Technologies in Tamil Computing, organized by the International Forum for Information Technology in Tamil. This gathering represents a significant milestone in the convergence of AI and the Tamil language, and its inclusion in our conference proceedings is a testament to the importance of this event. Tamil, a language of profound historical and cultural significance, is poised for a technological renaissance. AI stands as a pivotal force, shaping its future by facilitating innovation, enabling efficient communication, and preserving linguistic heritage. As the Principal of KCLASI am excited to witness the dynamic discourse that unfolds during this conference, which promisesto chart new frontiers in Tamil language computing.

The N.Mahalingam Tamil Research Centre, launched inthe enduring legacy of founder of Kumaraguru Institutions, Dr. N.Mahalingam Aiya, plays a pivotal role in emphasizing the importance of Tamil language and culture in the age of AI. Founded with a vision to advance research and scholarship in Tamil studies, the Center stands as a testament to the unwavering commitment to preserving and promoting the Tamil language in the digital era. The founder's enduring passion and dedication to Tamil studies serve as a guiding light for the institution, inspiring generations to delve into the language's rich history, literature, and traditions, and to harness the power of AI in language preservation and understanding. The significance of the "AI in Tamil Language Computing" topic lies at the intersection of preserving cultural heritage and embracing cutting-edge technology. Tamil, one of the world's oldest languages, boasts a rich literary tradition and an extensive global community. The application of AI in Tamil computing not only enables effective communication and information access for Tamil speakers but also safeguards and rejuvenates the language's historical and cultural value. The future of this topic promises groundbreaking advancements in natural language processing, machine learning, and data-driven language preservation. As AI continues to evolve, it will play an integral role in language revitalization, machine translation, and knowledge dissemination in Tamil, ultimately fostering global understanding and cooperation. The synergy between AI and Tamil computing presents an exciting frontier with far-reaching implications for linguistic diversity, technological innovation, and cross-cultural communication. It is an area ripe for exploration and investment, offering immense potential for enriching our digital world and preserving linguistic heritage for generations to come. This conference brings together experts, scholars, and practitioners in the fields of AI and Tamil computing for deep discussions, presentations, and shared experiences thatwill enrich our collective understanding of these intersecting domains. The conference will serve as a catalyst for groundbreaking research, collaborations, and the exchange of profound insights. The knowledge disseminated here will reverberate in the academic and technological realms for years to come.

Dr. Vijila Edwin Kennedy
Principal, KCLAS.

Greetings from Chairman, INFITT

அன்புள்ள அமைப்பாளர்கள், பங்கேற்பாளர்கள் மற்றும் விருந்தினர்களே! உலகத்தமிழ் தகவல் தொழில்நுட்ப மன்றம் (INFITT) மற்றும் அதன் உறுப்பினர்கள் சார்பாக, 2013 அக்டோபர் 13 அன்று கோயம்புத்தூரில் நடைபெற்ற "தமிழ்க் கணினியில் செயற்கை நுண்ணறிவு" சர்வதேச மாநாட்டை ஏற்பாடு செய்தவர்களுக்கு எனது மனமார்ந்த வாழ்த்துகளைத் தெரிவித்துக் கொள்கிறேன். INFITT மற்றும் குமரகுரு தொழில்நுட்பக் கல்லூரி இணைந்து ஏற்பாடு செய்த இந்த நிகழ்ச்சி, தமிழ் கணினி மற்றும் செயற்கை நுண்ணறிவுத் துறையில் குறிப்பிடத்தக்க மைல்கல்லை ஏற்படுத்தியுள்ளது. டிஜிட்டல் யுகத்தில் தொழில்நுட்பம் மற்றும் மொழியை மேம்படுத்துவதற்கான உங்கள் அர்ப்பணிப்பு உண்மையிலேயே பாராட்டுக்குரியது.

கல்வி மற்றும் தகவல் தொடர்பு முதல் வணிகம் மற்றும் அதற்கு அப்பால் பல்வேறு துறைகளில் செயற்கை நுண்ணறிவு மற்றும் தமிழ் கணினியின் ஒருங்கிணைப்பு புரட்சியை ஏற்படுத்தும் மகத்தான ஆற்றலைக் கொண்டுள்ளது. இந்தப் தளத்தில் புதுமை மற்றும் ஒத்துழைப்பை வளர்ப்பதில் உள்ள உங்களது முயற்சி உயர்ந்த பாராட்டுக்கு உரியவை. இந்த நிகழ்வின் போது தங்கள் அறிவையும் நுண்ணறிவையும் பங்களித்த அனைத்து பங்கேற்பாளர்கள், ஆராய்ச்சியாளர்கள் மற்றும் நிபுணர்களையும் நான் வாழ்த்த விரும்புகிறேன். உங்கள் மதிப்புமிக்க பங்களிப்புகள் சந்தேகத்திற்கு இடமின்றி விவாதங்களை செழுமைப்படுத்தி இந்தத் துறையில் மேலும் முன்னேற்றங்களுக்கு வழி வகுத்துள்ளது.

இந்த நிகழ்வின் வெற்றியைக் கொண்டாடும் வேளையில், தொழில்நுட்பம் மற்றும் மொழி ஆகியவற்றில் உள்ள வரம்பற்ற சாத்தியக்கூறுகளை ஆராய்வதில் தொடர்ந்து இணைந்து பணியாற்றுவோம். செயற்கை நுண்ணறிவுக்கும் தமிழ் கணினிக்கும் இடையே உள்ள இடைவெளியைக் குறைப்பதற்கான உங்கள் அர்ப்பணிப்பு, சந்தேகத்திற்கு இடமின்றி உலகளாவிய அளவில் நீடித்த தாக்கத்தை ஏற்படுத்தும்.

மீண்டும் ஒருமுறை, ஒரு வெற்றிகரமான நிகழ்வுக்கு வாழ்த்துகள், மேலும் இங்கு பகிரப்படும் அறிவும் கருத்துக்களும் வரும் ஆண்டுகளில் தொடர்ந்து முன்னேற்றத்தைத் தூண்டட்டும்.

அன்பான வாழ்த்துக்கள்,
இராம சுகந்தன்
தலைவர்
INFITT

Greetings from HOD, Tamil Department

இன்றைய காலத்தில் அறிவியல் புரட்சி நாளுக்கு நாள் மிகப்பெரிய அளவில் நடந்து கொண்டிருக்கிறது. இந்த அறிவியல் மாற்றத்திற்கேற்ப காலத்தோடு ஒரு மொழியும் பண்பாடும் தன்னைப் புதுபித்துக்கொள்ள வேண்டி இருக்கிறது. காலத்திற்கேற்ப தன்னை புதுப்பித்துக்கொள்ளும் மொழியே காலம் கடந்து நிற்கும். உலகில் உள்ள செம்மொழிகளில் ஒன்றான மூத்த தமிழ்மொழி இன்றளவும் ஒரு வாழும் மொழியாக இருக்கிறது. கல்வெட்டுக் காலத்தில் இருந்து தமிழ்மொழி தொடர்ந்து புதுப்பிக்கப்பட்டே வந்திருக்கிறது. இன்று செயற்கை நுண்ணறிவுத் தொழில்நுட்பம், தகவல் தொடர்பில் மிகப்பெரிய பிரளயத்தையே ஏற்படுத்தி வருகிறது.

இந்த செயற்கை நுண்ணறிவு செயல்பாடுகளும், பயன்பாடுகளும் தமிழ்மொழி தன் இயல்பான மாற்றங்களை உட்கொண்டு காலத்தால் அழியாத உலகில் தன்னை உறுதியான முறையில் நிலைநிறுத்திக்கொள்ளும். இவ்வகையில் இக்கருத்தரங்கம் மிகப்பெரிய பயனளிக்கும் என்று நம்புகிறேன். செயற்கை நுண்ணறிவுத்திறன் செயலிகளைத் தனக்கான, சகலவற்றிலும் வெளிப்படுத்திக்கொள்ளும் விதமாய் அறிஞர்கள் புதுமையாக்கினால் இன்னும் தமிழ்மொழி உலகத்தொடர்பில் மிக வலுவான உன்னத நிலையை அடையும் என்பதில் ஐயமில்லை.

தமிழைக் கணினிப்படுத்துவதில் ஆர்வம் காட்டி, முன்னேற்பாடுகள் செய்த குழுவிற்கும், கற்கும் ஆர்வம் கொண்டு பங்கேற்க வந்த அனைத்து நல்லுங்களுக்கும் என் மனப்பூர்வமான வாழ்த்துகள்.

முனைவர். வேணுகோபால் S,
தலைவர், தமிழ்த்துறை,
குமரகுரு பன்முகக் கலை அறிவியல் கல்லூரி.

Chief Guest

Dr Arun Janarthanan,
Technology Director and Engineering Practice Head, HCL Technology, Chennai.

Speakers

1. **Dr S.K. Lavanya**,
   Anna Univ MIT Campus.
   Topic: Building Transformer Based Models for NLP Applications

2. **Dr C. N. Subalalitha**,
   SRM Inst of Science & Technology, Chennai.
   Topic: Natural Language Processing, Machine Learning, Discourse Analysis and Computational Linguistics

3. **Prof. D. Thenmozhi**,
   SSN College of Engineering, Chennai.
   Topic: Information Extraction from Tamil Medicinal Documents

4. **Dr K. Parameswari**,
   Indian Inst. Of Information Technology, Hyderabad.
   Topic: Advances in Tamil Computing: from shallow parsing to Machine Translation

5. **Dr T. Uthayasankar**,
   Univ of Moratuwa, Sri Lanka.
   Topic: Impact of LLMs in the World

6. **Dr R. Ponnusamy**,
   Chennai Inst. Of Technology, Chennai.
   Topic: Tamil Text Generation using GPT-3 Models

7. **Prof. Vasu Renganathan**,
   Univ of Pennsylvania, PA, USA.
   Topic: Large Language Models and the Role of Linguists in the World of AI

8. **Dr R Selvajyothi**,
   Univ of Malaya, Malalysia.
   Topic: Modern Scientific Research Approaches in Tholkaappiam

9. **Thirukkural Ganesan**,

Researcher, Coimbatore.
Topic: Artificial Intelligence in Thirukkural

10. **Mr Raju Kandasamy**,
Principal Consultant, Thoughtworks, Coimbatore.
Topic: From Alphabets to AI - Building Tamil GPT

11. **Mr Rajaram @ Neechalkaran**
Infosys
Topic: Tools for Tamil Computing

Organizers

1. **Mr Mahendran Arjun Raja**,
Mentor, Kumaraguru Tamil Mandram
LOC Chair - INFITT Conference

2. **Dr Jagadeesan R**,
Director,
N Mahalingam Tamil Research Centre (NMTRC)

3. **Dr Sivakumar S**,
Associate Dean,
School of Foundational Sciences.

4. **Prof. Sri Geetha M**,
Assistant Professor,
Department of Artificial Intelligence and Data Science

5. **Prof. Nithya S M**,
Assistant Professor,
Department of Information Technology

6. **Prof. Kirubakaran R**,
Assistant Professor,
Department of Computer Science and Engineering

7. **Dr Sunil Joghee G**,
Assistant Professor, Tamil Department, KCLAS &
Staff Coordinator - Kumaraguru Tamil Mandram

8. **Prof. Sudalaimani P**,
Assistant Professor, Tamil Department, KCLAS & Staff Coordinator - Kumaraguru Nithilam

9. **Mr Nivethan A M**,

Senior Manager,
Kumaraguru Institutions

10. **Mr Nethaji Subash M,**
    Senior Executive - Sustainability & Alumni, Kumaraguru Tamil Mandram & Nithilam

Student Coordinators

1. **Goutham T**,
   Second year, Master of Business Administration.
   Advisor, Kumaraguru Tamil Mandram and Nithilam

2. **Janaranjani R**,
   Final year, Electrical and Electronics Engineering
   President, Kumaraguru Tamil Mandram

3. **Inbavel T**,
   Final Year, Electronics and Communications Engineering
   President, Kumaraguru Nithilam.

4. **Allwin Kenneth P**,
   Final Year, Computer Science and Engineering
   President, Department Association - Computer Science and Engineering

5. **Kavin M**,
   Final year, Civil Engineering.
   Vice President, Pixel'D

6. **Krithik Sivasubramanian**,
   Final year, Mechanical Engineering
   PRO, Studio KCT

7. **Janarthanan S**,
   Final Year, Electronics and Communications Engineering

8. **Surya M**,
   Final Year, Electronics and Communications Engineering

9. **Yuvaraja M**,
   Final year, Civil Engineering
   Advisor, NSS, RRC, VBC

10. **Karthikeyan S**,
    Final year, Civil Engineering
    Head, Communications, Department Association

11. **Karthikeyan M P**,
    Final Year, Information Technology

12. **Shrivathsan G**,
    Final Year, Information Technology

Student Coordinators

13. **Kaviya Priya M**,
    Third year, Artificial Intelligence and Data Science
    Secretary, Kumaraguru Tamil Mandram

14. **Siddhartha Devan V**,
    Third year, Artificial Intelligence and Data Science
    Secretary, Kumaraguru Nithilam.

15. **Balaji R**,
    Third year, Artificial Intelligence and Data Science
    Treasurer, Kumaraguru Nithilam

16. **Jumaytha Thabaseen S**,
    Third year, Bachelor of Business Administration
    Former President, KCLAS Tamil Mandram

17. **Ragul Adhithya M,**
    Third year, Information Technology
    President, Land of Lexicons

18. **Mukilarasu R**,
    Third Year, Artificial Intelligence and Data Science
    Executive Member, Studio KCT

19. **Pravin P B**,
    Third Year, Artificial Intelligence and Data Science
    Secretary, Pixel'D

20. **Kavin Bharathi R M**,
    Third Year, Artificial Intelligence and Data Science
    Research and Innovation Coordinator, Department Association – AI&DS

21. **Vettrikanth S**,
    Third Year, Artificial Intelligence and Data Science
    Media and Marketing Coordinator, Department Association – AI&DS

Student Coordinators

22. **Pranavi K R**,
    Second year, Political Science
    President, KCLAS Tamil Mandram

23. **Kavishree S**,
    Second Year, Electrical and Electronics Engineering
    Joint Secretary, Kumaraguru Nithilam

24. **Aswath Alifa M**,
    Second Year, Political Science

25. **Kavin Prasath M G**,
    Second Year, Biotechnology

26. **Bavatharani G**,
    First Year, Tamil Creative Writing

International Conference on Tamil Computing

Invited Talks

Advancements in Tamil Computing: Shallow Parsing to Machine Translation

Parameswari Krishnamurthy
Language Technology Research Centre
International Institute of Information Technology- Hyderabad
<param.krishna@iiit.ac.in>

Abstract

This paper provides a comprehensive overview of the evolving landscape of Tamil computing, spanning from the foundational concept of shallow parsing to cutting-edge developments in machine translation using different techniques spanning from Rule-based to Machine Learning approaches. Tamil, with its rich linguistic nuances and intricate morphology, presents unique challenges and opportunities in the realm of computational linguistics. Through this journey, we will explore how shallow parsing techniques have been instrumental in understanding and processing Tamil text, and then delve into the transformative potential of machine translation for bridging language barriers and promoting cross-cultural communication.

1. Introduction:

Tamil boasts a rich literary and cultural heritage. Over the years, advancements in computing have played a pivotal role in the development and enhancement of Tamil language processing. From shallow parsing to machine translation, the journey of Tamil computing has been marked by remarkable progress. This paper discusses the key requirements in Tamil computing, emphasizing the need for shallow parsing techniques to contemporary machine translation systems. There have been numerous endeavors dedicated to developing technology tools and applications for the Tamil language (Rajendran, S., et.al, 2018). Several educational institutions, independent researchers, Tamil enthusiasts, as well as national and international companies have been dedicating years of effort to develop technology for the Tamil language. There is a great significance of shallow parsing in agglutinative languages like Tamil which has a unique morphological structure where words are constructed by combining multiple morphemes or affixes. Understanding and extracting these morphemes are pivotal for various language processing tasks.

2. Shallow Parsing in Tamil Language Processing

Shallow parsing, is an initial step in language processing, involving the identification of tokens, parts-of-speech and finding out the phrasal units or "chunks" within sentences. In the context of Tamil computing, early efforts were dedicated to developing shallow parsing algorithms to extract meaningful linguistic units like noun phrases, verb phrases, etc., from Tamil text.

Researchers employed rule-based approaches and linguistic heuristics to achieve this, creating the foundational groundwork for more sophisticated language processing techniques.

2.1. Tokenization and Multi-Token word segmentation:

Tokenization in Tamil, as in any language, involves breaking a text into smaller units called tokens. These tokens can be fully formed words, subwords, characters, or phrases, depending on the granularity of analysis. Spaces within written language play a critical role in token identification. However, in Tamil, there are frequently occurring multi-token words (MTW) that necessitate segmentation for subsequent processing steps. Examples:

Determinar+noun: இப்பகுதி [‘இப் (this)’, ‘பகுத (part)’]
noun+verb: இடமாகும் [‘இடம் (place)’, ஆகும் (is)’]

MTW tends to add multiple grammatical pieces of information within a word and cannot be identified with a single POS. A token status is required for further analysis.

2.2. Morphological Analysis:

Morphological analysis is a crucial component in Tamil language processing. Morphological analysis involves breaking down words into morphemes, aiding in understanding word structures and inflections in Tamil. Inflections, being modifications or additions to a word, impart crucial grammatical and contextual information. For instance, in Tamil, verb conjugations, tense markers, case markers, and gender agreements are often expressed through inflections. By recognizing and analyzing these inflections, we can unravel the syntactic and semantic nuances embedded within the language, enabling more accurate and insightful language processing. Using universal dependency morph features, the following features are crucial which are realized as inflections within word in Tamil.

Type	Features	Values
Pronominal type	PronType	Personal (Prs), Reflexive (Rfl), Reciprocal (Rcp), Interrogative (Int), Relative (Rel), Indefinite (Ind)
Numeral Type	NumType	Cardinal (Card), Ordinal (Ord), Fraction (Frac)
possessive	Poss	Yes
reflexive	Reflex	Yes
foreign word	Foreign	Yes
abbreviation	Abbr	Yes

Wrong spelling	Typo	Yes
gender	Gender	Masculine (Masc), Feminine (Fem), Neuter (Neut)
animacy	Animacy	Animate (Anim), Human (Hum), Inanimate (Inan)
Number	Number	Singular (Sing), Plural (Plur)
Case	Case	Nominative (Nom), Accusative (Acc), Instrumental (Ins), Dative (Dat), Ablative (Abl), Allative (All), Benefactive (Ben), Comitive (Com), Locative (Loc), Genitive (Gen), Vocative (Voc), other Oblique cases (Obl)
definiteness	Definite	Definite (Def), Indefinite (Ind)
Verbal form	VerbForm	Finite (Fin), Infinite (Inf), Participle (Part), Gerund (Ger), Converb (Conv)
mood	Mood	Indicative (Ind), Imperative (Imp), Conditional (Cnd), Potential (Pot), Desiderative (Des), Necessity (Nec)
tense	Tense	Present (Pres), Past (Past), Future (Fut)
aspect	Aspect	Progressive (Prog), Perfective (Perf),
voice	Voice	Active (Act), Passive (Pass), Causative (Cau)
polarity	Polarity	Positive (Pos), Negative (Neg)
person	Person	1, 2, 3
polite	Polite	Formal (Form)
clusivity	Clusivity	Inclusive (In), Exclusive (Ex)

Table 1: Morph Features in Tamil

We aim to build a context-aware morphological analysis that involves analyzing the structure of words in a sentence while considering the surrounding context. Table-2, the row "morph" provides the example features on context-sensitive morphological analysis.

2.3. Part-of-Speech Tagging:

Part-of-speech (POS) tagging assigns grammatical categories (e.g., noun, verb, adjective) to words, facilitating further analysis and understanding of sentence structure. These analyses serve as essential building blocks for subsequent natural language processing tasks. We have used the following tags while annotating Tamil texts with POS categories:

Open class categories: ADJ (adjective), ADV (adverb), INTJ (interjection), NOUN (noun), PRON (pronoun), VERB (verb),

Closed class categories: ADP (adposition), AUX (auxiliary), CCONJ (coordinating conjunction), DET (determiner), NUM (numeral), PART (particle), PROPN (proper noun), SCONJ (subordinating conjunction),

Other: SYM (symbol), PUNCT (punctuation)

The table 2, the row "pos" provides the example of POS information of tokens.

2.4. Chunking:

Chunking involved grouping smaller units or constituents in a sentence to understand their phrasal structure. This method helps to identify phrases and their relationships, providing valuable insights into syntax and language comprehension. The example chunk output is given here:

(1) (*நான்*)_NP (*அந்த பபண்ணை*)_NP (*குதணையுடன்*)_NP (*பார்த்ததன்*)VM.

3. Syntactic Parsing for Tamil:

Building a parser is a challenging task as it deals with structural ambiguities in languages. Structural ambiguities are realized due to two different factors; (i) attachment ambiguity and (ii) coordination ambiguity. They are illustrated in examples (2) and (3) respectively:

(2) *நான் அந்த பபண்ணை குதணையுடன் பார்த்ததன்.*

`I saw the girl with a horse'

(3) *வயதான ஆண்களும் பபண்களும் வந்தார்கள்.*
'Old men and women came'

In the example (2), the attachment of the associative case marked noun phrase (NP) (the horse) shows an ambiguity as it can be potentially attached/associated with the subject NP (nāṉ 'I') or with the object NP (peṇ 'the girl'). In example (3), the coordination ambiguity is shown where the adjective 'old' may have been interpreted to be coordinated with either 'men' or 'women'. The parser attempts to resolve such structural ambiguities based on various factors such as morphological, syntactic, semantic, contextual and discourse knowledge of a language. Once ambiguities are identified, the parser attempts to choose correctly parsed output with the given knowledge. The annotated treebank of Tamil data is given in Table (2) with colums token number (tkn_no), token(tkn), parts-of-speech (pos), morph features(morph), relation token (rel_tkn) and syntactic relation (syn_rel).

#text= ராமனின மனைவி சீதையை ராவணன இலங்கைக்குக கடதத்ச சனறான.

tkn_no	tkn	pos	morph	rel_tkn	syn_rel
1	ராமனின்	PROPN	Case=Gen|Gend=Masc|Number=Sing	2	nmod
2	மனைவி	NOUN	Case=Nom|Gend=Fem|Number=Sing	3	appos
3	சீதையை	PROPN	Case=Acc|Gend=Fem|Number=Sing	6	obj
4	ராவணன்	PROPN	Case=Nom|Gend=Masc|Number=Sing	6	nsubj
5	இலங்கைக்குக்	PROPN	Case=Dat|Gend=Neu|Number=Sing	6	obl:to
6	கடத்திச்	VERB	Polarity=Pos|VerbForm=Conv	0	root
7	சென்றான்	AUX	Gender=Masc|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	6	aux
8	.	PUNCT	PunctType=Peri	6	punct

Table 2: Example for Tamil Syntactic Treebank

4. Implementation methodology of Shallow Parsing and Syntactic Parsing

There are lots of off-the-shelf machine learning algorithms available in training parsers. One of the recent algorithms which reported with reasonale accuracy is Trankit (Van Nguyen et al, 2021). Initially, we would like to explore this algorithm and customize it to the need of Tamil parsing. Trankit is a lightweight Transformer-based Toolkit for multilingual Natural Language Processing. It delivers a trainable pipeline for fundamental NLP tasks for over 100 languages and 90 pretrained pipelines for 56 languages. The Trankit toolkit outperforms prior multilingual NLP pipelines over sentence segmentation, part-of-speech tagging, morphological features tagging, and dependency parsing while holding competitive performance for tokenization, multi-word token expansion, and

lemmatization. The memory usage and speed are efficient despite using large pretrained transformer models.

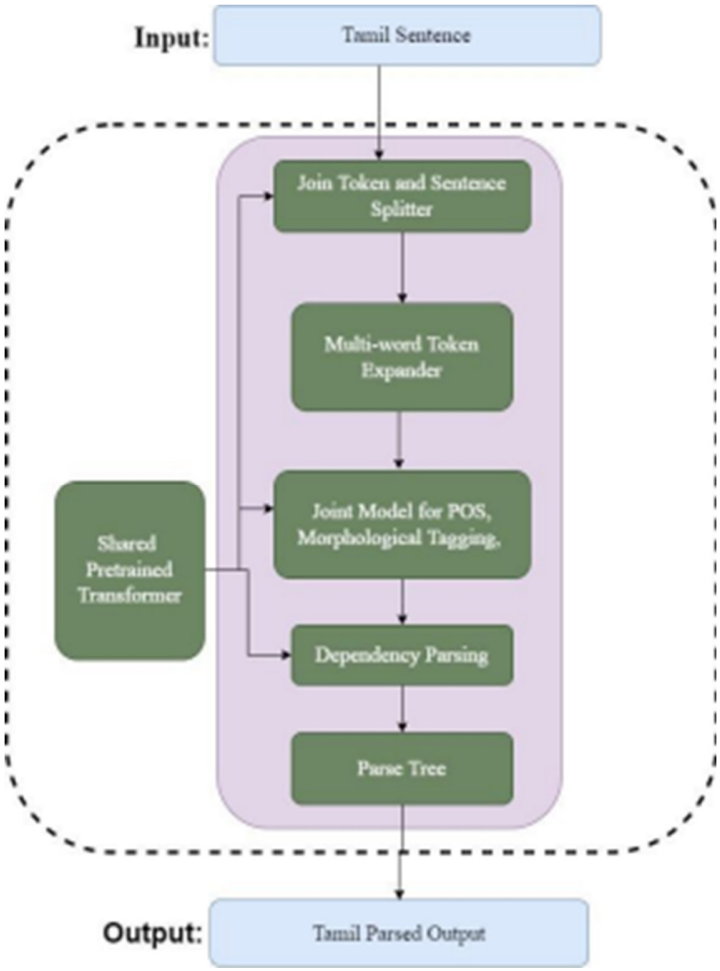


Figure-1: Architecture of Parsing

5. Importance of Parsers in Neural Machine Translation involving Tamil:

Machine translation involves translating text from one language to another using computational algorithms. In the context of Tamil, machine translation has seen significant advancements in recent years. Early machine translation systems relied on rule-based approaches and linguistic knowledge. However, with the advent of neural machine translation (NMT), Tamil translation systems have become more accurate and contextually relevant. NMT leverages deep learning models to translate sentences and has drastically improved the quality of translation in both written and spoken Tamil.

Parsers play a crucial role in Neural Machine Translation (NMT) by aiding in the understanding and analysis of input sentences and their subsequent conversion into target language output.

Here are some key points highlighting the importance of parsers in NMT:

Syntactic Understanding: Parsers help in analyzing the syntax and structure of sentences, identifying relationships between words and phrases. This understanding is essential for generating coherent and grammatically accurate translations.
Semantic Representation: Parsers assist in capturing the semantic meaning of the input sentence. Understanding the meaning of the sentence allows the NMT system to produce more contextually relevant and accurate translations.

Improved Word Alignment: Accurate parsing helps in aligning words and phrases between the source and target languages. Proper word alignment is critical for generating precise translations, ensuring that words are appropriately matched based on their meanings and positions.

Phrase Extraction and Chunking: Parsers facilitate the extraction of phrases and chunks from the input sentence. These chunks can be translated more effectively, enabling the system to generate natural and fluent translations.

Handling Ambiguity: Natural languages often have ambiguous sentence structures. Parsers help in disambiguating such sentences by analyzing the context and providing the most probable syntactic and semantic interpretations. This is particularly vital for generating accurate translations in the presence of ambiguities.

Error Detection and Correction: Parsers aid in identifying errors in the input sentence, such as grammatical mistakes or inconsistencies. This allows the NMT system to produce more refined translations by detecting and correcting errors in the source text.

Incorporating parsers within the NMT framework enhances the system's ability to comprehend the input text at a deeper level. This deeper comprehension, facilitated by syntactic and semantic analysis, significantly contributes to the production of higher-quality and more accurate translations in neural machine translation.

6. Conclusion:

Advancements in Tamil computing, from shallow parsing to modern machine translation, have transformed the landscape of Tamil language processing. The progress made in shallow parsing, including morphological analysis, part-of-speech tagging, chunking and syntactic parsing, and semantic parsing has laid the foundation for robust language understanding. Furthermore, the introduction of neural machine translation has significantly enhanced the accuracy and fluency of translating Tamil text, enabling better communication and collaboration in a globalized world. Within Neural Machine Translation (NMT), parsers have a significant impact, contributing to the interpretation and analysis of input sentences, thus influencing the quality of the resulting translation in the desired target language. For Tamil, a language renowned for its morphological complexity, leveraging shallow parsing and syntactic parsing are not merely a choice but a necessity to unlock the true potential of language technology applications and enhance communication in the digital realm.

References

Akoury, N., Krishna, K. and Iyyer, M., 2019. *Syntactically supervised transformers for faster neural machine translation*. arXiv preprint arXiv:1906.02780.

Bharati, A., Sharma, D. M., Husain, S., Bai, L., Begam, R., and Sangal, R. 2009. Anncorra: Treebanks for Indian Languages, Guidelines for Annotating Hindi Treebank. Retrieved on 1st December, 2016. http://aclweb.org/anthology/W17-63

Bunt, H., Carroll, J. and Satta, G. eds., 2005. *New developments in parsing technology* (Vol. 23). Springer Science & Business Media.DOI- https://doi.org/10.1007/1-4020-2295-6

Clark, A., Fox, C. and Lappin, S. eds., 2012. *The handbook of computational linguistics and natural language processing* (Vol. 118). John Wiley & Sons.

Comrie, Bernard. 1991. 'Form and Function in Identifying Cases'. In Frans lank (ed.),

De Marneffe, M.C., Manning, C.D., Nivre, J. and Zeman, D., 2021. Universal dependencies. *Computational linguistics*, *47*(2), pp.255-308.

Dhivya, R., Dhanalakshmi, V., Anand Kumar, M. and Soman, K.P., 2011, December. Clause boundary identification for tamil language using dependency parsing. In International Joint Conference on Advances in Signal Processing and Information Technology (pp. 195-197). Springer, Berlin, Heidelberg. Institute of Linguistics and Culture.

Eriguchi, A., Tsuruoka, Y. and Cho, K., 2017. *Learning to parse and translate improves neural machine translation*. arXiv preprint arXiv:1702.03525.

Kulkarni, A. 2019. Sanskrit Parsing Based on the Theories of śābdabodha. Indian Institute of Advanced Study, Shimla and D. K. Publishers(P) Ltd.

Lehmann, Thomas. 1993. A Grammar of Modern Tamil. Pondicherry: Pondicherry

Nivre, J. 2006. Inductive Dependency Parsing. In T*ext, Speech and Language Technology. Netherlands*: Springer. Vol 34.

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N. and Tsarfaty, R., 2016, May. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).

Parameswari K. and Sarveswaran, K., 2022. Towards Building a Modern Written Tamil Treebank. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)* (pp. 61-68).
Pu, D. and Sima'an, K., 2022, June. Passing parser uncertainty to the transformer: Labeled dependency distributions for neural machine translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation (pp. 41-50).

Rajendran S, R., Anandkumar, M., Dhanalakshmi, V. and SN, M.R. 2019. A Parser for Question answer System for Tamil. 2019, Conference Papers, 18th Tamil Internet Conference.

Ramasamy, L., and Žabokrtský, Z. 2011. Tamil Dependency Parsing: Results Using RuleBased And Corpus Based Approaches. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Berlin: Springer, pp. 82-95.

Sarveswaran, K. and Dias, G., 2020. ThamizhiUDp: A dependency parser for Tamil. *arXiv preprint arXiv:2012.13436*.

Straka, M. and Straková, J., 2017, August. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88-99).

Sureka, K., Srinivasagan, K. G., and Suganthi, S. 2014. An Efficiency Dependency Parser Using Hybrid Approach for Tamil Language. Retrieved on 20th December, 2017. https://arxiv.org/abs/1403.6381.

Tesnière, Lucien, ed. 1959. Éléments de Syntaxe Structurale. Klinck-sieck Paris.

Vijay Sundar Ram and Sobha L. 2021. *Dependency Parsing in a Morphological rich language, Tamil.* In proceedings of Workshop on Parsing and its Applications for Indian Languages (PAIL), co-located with ICON 2021, 16 December 16, 2021. ACL Anthology.

Van Nguyen, M., Lai, V.D., Veyseh, A.P.B. and Nguyen, T.H., 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. *arXiv preprint arXiv:2101.03289*.

Generative AI for the Tamil Language

Dr. Subalalitha C.N
Associate Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Chennai-603203 &
Founder and Director, First Language Technologies Private Limited, Chennai

1.Introduction to Generative AI for Tamil Language

Generative Artificial Intelligence (AI) has made remarkable strides in natural language processing and generation, enabling machines to understand, generate, and manipulate human languages. One of the exciting applications of generative AI is its ability to work with languages beyond the commonly spoken ones, such as English. In this two-page summary, we will explore the significance and developments of generative AI in the context of the Tamil language.

Tamil is one of the oldest and classical languages in the world, with a rich literary heritage. It is predominantly spoken in the Indian state of Tamil Nadu and Sri Lanka, with a significant diaspora worldwide. The utilization of generative AI for Tamil holds immense potential for a variety of applications, including content generation, language translation, chatbots, and more.

2.Development of Tamil Language Models

The foundation of generative AI for Tamil begins with the creation of robust language models, similar to models developed for English. Large Language Models like GPT, BARD have made significant strides in advancing the capabilities of AI models for various languages, including Tamil. Training language models on vast corpora of Tamil text, encompassing both contemporary and classical literature, has allowed these models to understand and generate Tamil text with impressive fluency and coherence.

3.Applications of Generative AI for Tamil Language

Language Translation

1.: Generative AI has greatly facilitated Tamil-English and English Tamil translation. These models can produce highly accurate translations, aiding in cross lingual communication, content localization, and accessibility.

Content Generation

2.: Content creation, whether for articles, marketing materials, or creative writing, benefits from generative AI. AI-generated content in Tamil can be a boon for businesses and individuals seeking to produce content efficiently.

Chatbots and Virtual Assistants

3.: Conversational AI powered by generative models can provide intelligent and context-aware responses in Tamil. This is valuable for customer support, virtual assistants, and other interactive applications.

Language Learning

4.: Generative AI can assist in language learning by generating exercises, quizzes, and practice sentences in Tamil, making the learning process more engaging and personalized.

Text Summarization

5.: Summarizing lengthy Tamil texts or news articles can be automated using generative models, saving time and effort for readers and researchers.

4. Challenges and Considerations

While the development of generative AI for the Tamil language is promising, it also comes with several challenges:

Data Availability

1.: Access to large and diverse datasets is crucial for training effective language models. For Tamil, ensuring a sufficient amount of high-quality text data is a challenge.

Bias and Fairness

2.: As with any AI system, addressing bias and ensuring fairness in generative AI for Tamil is essential to avoid perpetuating stereotypes or promoting harmful content.

Resource Intensity

3.: Training and deploying generative AI models can be computationally intensive, which may limit accessibility for smaller organizations or individuals.

Evaluation Metrics

4.: Developing appropriate evaluation metrics for Tamil language models is essential to measure their effectiveness accurately.

5. Conclusion

Generative AI for the Tamil language is a significant step forward in harnessing the power of artificial intelligence to enhance various aspects of Tamil communication, education, and content creation. As AI research and development continue to advance, we can expect even more sophisticated and tailored solutions for the Tamil language, further promoting its preservation and growth in the digital age.

6. Future Prospects and Recommendations

The future of generative AI for the Tamil language is promising, and there are several areas where further development and attention are needed:

Data Collection and Curation

1.: Efforts should be made to collect and curate diverse and high-quality datasets in Tamil, spanning different domains and registers, to improve the training and performance of language models.

Community Engagement

2.: Collaboration with the Tamil-speaking community, including linguists, writers, and educators, is essential in shaping the development of AI systems for Tamil. Their expertise can ensure culturally sensitive and accurate language generation.

Ethical Guidelines

3.: Establishing ethical guidelines for the use of generative AI in Tamil is vital. Transparency, fairness, and accountability should be at the forefront of AI development.

Accessibility

4.: Making generative AI tools and applications for Tamil accessible to a wider audience, including those with limited technical expertise, should be a priority.

Multimodal Integration

5. : Expanding generative AI to support not only text but also speech and visual inputs can enhance its utility in various applications, such as video subtitles, voice assistants, and image captioning.

In conclusion, generative AI for the Tamil language has the potential to revolutionize how Tamil content is generated, translated, and interacted with in the digital age. While challenges exist, continued research, collaboration, and community involvement can lead to the development of powerful and ethically sound AI systems that benefit the Tamil-speaking population and contribute to the broader field of natural language processing.

Large Language Models (LLM) and the Role of linguists in the World of AI

Vasu Renganathan
University of Pennsylvania
vasur@sas.upenn.edu
(http://www.sas.upenn.edu/~vasur/project.html)

Introduction:

Large Language Models (LLMs) and correspondingly building vector databases in the recent time produced appealing advances in enabling machines to behave like humans. Linguists can play a vital role in training LLMs by using many linguistic theories such as lexical semantics, semiotics and so on to identify and account for complexities in meanings of words and expressions so a much more robust use of this research can be achieved. For example, the Tamil word "paṭi" has multiple meanings, such as "to read", "to settle", “step”, “a measuring container” and so on. Not only the homonymous words but also extended meanings of expressions as in vayiṟu eṟikiṟatu to mean both ‘burning sensation in stomach’ as well as ‘feeling hurt mentally due to unusual attitudes of others’ would also pose problems for machines in parallel to how humans understand the natural language. A linguist could help an LLM to be trained with such nuances of expressions so a live interpretation can be achieved by machines. Such efforts would allow the machine to interpret dialogues in a more meaningful manner than now. Prediction and probabilities within complex linguistic structures need to be accounted for more comprehensively than before in order to design plausible and life-like machine-human interactions. Linguistic theories have traditionally focused more on structures than on the bidirectional predictability of words or even sentences, which is what systems like BERT and GPT attempt to do (Cf. Noir 2020). An attempt is made here to demonstrate using a robot as to how the recent developments of NLP can be implemented and tested for Tamil.

Recent Advances in NLP tasks and development of a Robot:

With an extensive research and plausible outcome from many open-source projects such as text to speech, machine translation, Wiki-resources and speech to text, it is now quite possible to integrate them into a mini-robot type of machines and converse to them in a natural way. This project aims at a similar effort with a robot and attempts to converse in Tamil with it. The activities include commanding the robot to move around such as forward, backward, circle around and so on with commands in Tamil, like முன்னால போங்க, பின்னால வாங்க, சுத்துங்க, respectively. This system also can be trained with particular set of movements and use later to perform them using commands in Tamil. For example, navigating from one place to another can be recorded in sequence and link it to commands such as சமையலறைக்கு போங்க, சமையலறைக்கு போயிட்டு வாங்க and so on so forth, so these commands will call the routine that was trained earlier. The robot that was built earlier and is demonstrated in

http://robot.tamilnlp.com extensively uses a voice recognition card called EasyVR and its speech to text capability is restricted to a single person. Whereas, the current project employs Google's text to speech and speech to text APIs extensively and attempts to process Tamil voice in a natural way independent of the speaker. The issue with this project, however, lies in the way the parsing of Tamil sentences is done in a meaningful manner, so the machine can respond with its text to speech API. What is crucial is the processing task of the text obtained between the two processes namely speech to text and text to speech. The speech to text API attempts to convert most of the spoken Tamil expressions into corresponding literary text. What is of utmost importance is that one needs to find the input and output of this API with a set of data and analyze the commonalities, so the processing task can be done efficiently. For example, the output நீங்கள் போய்க்கொண்டு இருக்கிறீர்களா, நீங்கள் நாளைக்கு என்னுடைய வீட்டுக்கு உங்களுடைய புத்தகத்தை எடுத்துக் கொண்டு வருவீர்களா are interpreted from the corresponding literary form of the speech. However, the spoken form of the output நீங்க நாளைக்கு என்னோட வீட்டுக்கு உங்க புத்தகத்தை எடுத்துட்டு வருவீங்களா requires special attention. The word நாளைக்கி is understood/interpreted as நாளைக்கு, எடுத்துக்கிட்டு is interpreted as எடுத்துட்டு. When these words are uttered separately, படிச்சுக்கிட்டு, எடுத்துக்கிட்டு are produced as output with "க்க்" in them. Strangely, when these words are included as part of a sentence the output produces only எடுத்துட்டு as in உங்க புத்தகத்தை எடுத்துட்டு வருவீங்களா. Obviously, this kind of analysis needs to be done with a large set of expressions and a common feature of the output from this application needs to be arrived at, so further processing of the output from speech to text API can be possible. Particularly, this kind of study is needed for applications like question/answering systems, database retrieval systems and so on.

Using the Wiki resources in Tamil:

When interacting with the Wiki resources, an attempt is made to parse sentences with the word "பத்தி' and the API to Wiki resources is called. Following are some of the interactions and results from this device:

Input: ஒரு பத்தி சொல்லுங்க:

Output: ஒரு கல் ஒரு கண்ணாடி, உதயநிதி ஸ்டாலின் நடிக்க ராஜேஷ் இயக்கத்தில் வெளிவந்த தமிழ் காதல் - நகைச்சுவைத் திரைப்படம். இத்திரைப்படத்தில் தயாரிப்பாளர், உதயநிதி ஸ்டாலின், கதாநாயகனாக அறிமுகமானார்.

Input: குமரகுரு கல்லூரி பத்தி சொல்லுங்க

Output: குமரகுரு தொழில்நுட்பக் கல்லூரி கோவையில் அமைந்துள்ள ஒரு தனியார் பொறியியல் கல்லூரி. 1984 ஆம் ஆண்டு தொடங்கப்பட்ட இக்கல்லூரி சத்தி வணிக குழுமத்தின் இராமானந்த அடிகளார் அறக்கட்டளையினால் நிருவகிக்கப்படுகிறது.

Although the second query and result seem to be desired, the first query and result requires attention as the text from the query and result do not have any direct link in any of the phrases within the output. However, the input ஒரு கல் கண்ணாடி பத்தி சொல்லுங்க does indeed produce the same output as above. Except for such minor issues, accessing API resources from Tamil Wiki offers a good result for the names

of towns, institutions, projects and so on.

Input: மயிலாடுதுறை பற்றி சொல்லுங்க:

Output: மயிலாடுதுறை (Mayiladuthurai) (முன்பு மாயவரம் என்று அழைக்கப்பட்டது) இந்தியாவில், தமிழ்நாடு மாநிலத்தில் மயிலாடுதுறை மாவட்டத்தில் உள்ள நிர்வாகத் தலைமையிடமும், சிறப்பு நிலை நகராட்சியும் ஆகும். மயில்கள் ஆடும் துறை என்பதால் மயிலாடுதுறை என அழைக்கப்படுகிறது.

Input: மதுரை திட்டம் பற்றி சொல்லுங்க:
மதுரை தமிழ் இலக்கிய மின் தொகுப்புத் திட்டம் (Project Madurai) என்பது தமிழ் இலக்கியங்களை, இணையத்தில் இலவசமாக வெளியிடும், ஒரு திறந்த, தன்னார்வ, உலகளாவிய முயற்சியாகும். 1998 ஆம் ஆண்டு, பொங்கல் தினத்தன்று, தொடங்கப்பட்ட இத்திட்டம், இன்றளவில் தொடர்ந்து இயங்கி வருகிறது.

Translation resources:

Google's translation resources are incorporated in this device and the commands such as ஹிந்திலெ சொல்லுங்க, மலயாளத்துல சொல்லுங்க, கன்னடால சொல்லுங்க etc., followed by any sentence in Tamil is translated to the corresponding target languages. Following are some of the outcome from this device:

ஹிந்தில சொல்லுங்க
हिंदी में कहें
hindee mein kahen
நீங்களும் உங்க தம்பியும் நாளைக்கு என்னோட வீட்டுக்கு சாப்பிட வரீங்களா
क्या तुम और तुम्हारा भाई कल रात क□ खान□क□ लिए मर□ घर आओग□
kya tum aur tumhaara bhaee kal raat ke khaane ke lie mere ghar aaoge?
தமிழ் நல்லா □ச வருமா
क्या आप □च्छी तमिल बोल सकत□हैं?
kya aap achchhee tamil bol sakate hain?
நீங்க எல்லாரும் இந்தியாவுல எந்தெந்த இடத்துக்கு □ாய் இருக்கீங்கன்னு ஒன்னு விடாம சொல்றீங்களா
क्या आप मुझ□बता सकत□हैं कि आप सभी भारत में कहां ग□□हैं?
kya aap mujhe bata sakate hain ki aap sabhee bhaarat mein kahaan gaye hain?

கனடால சொல்லுங்க
ಕೆನಡಿಯನ್ ಎಂದು ಹೇಳಿ
Kenaḍiyan endu hēḷi
நீங்க நாளைக்கு என்னோட வீட்டுக்கு வரீங்களா
ನಾಳೆ ನನ್ನ ಮನೆಗೆ ಬರುತ್ತೀಯಾ?
Nāḷe nanna manege baruttīyā?
நீங்க காலையில சாப்பாடு சாப்பிட்டீங்களா
ನೀವು ಬೆಳಿಗ್ಗೆ ತಿಂದಿದ್ದೀರಾ?

Nīvu beḷigge tindiddīrā?

The “speech to text” system converts input voice into Tamil text and the translation API translates the respective text to corresponding target language and outputs in the respective script. Correspondingly, the output text is both recorded in the database and converted to corresponding voice using the “text to speech” application. Each of the target language text as given here are uttered by the text to speech engine API correspondingly by exploiting all of the resources.

Accessing online text databases:

It’s quite possible to access online literary databases and convert them to voice as well as text with these resources. Following example shows how this device is used to access online Thirukkural databases:

Input: குறள் 52:
Output in both text as well as in voice: 52. பொருளாட்சி போற்றாதார்க்கு இல்லை அருளாட்சி ஆங்கில்லை ஊன்தின் பவர்க்கு. பொருள் ஆட்சி போற்றார்தார்க்கு இல்லைபொருளால் பயன் கோடல் அதனைப் பாதுகாவாதார்க்கு இல்லை; ஆங்கு அருள் ஆட்சி ஊன் தின்பவர்களுக்கு இல்லைஅது போல அருளாள் பயன் கோடல் ஊன் தின்பவர்களுக்கு இல்லை. விளக்கம் (பொருட்பயன் இழத்தற்குக் காரணம் காவாை□ போல, அருட்பயன் இழத்தற்கு ஊன் தின்னல் காரணம் என்பதாயிற்று. ஊன் தின்றாராயினும் உயிர்கட்கு ஒரு தீங்கும் நினையாதார்க்கு அருள் ஆள்தற்கு இழுக்கு இல்லை என்பாரை □றுத்து, அஃது உண்டு என்பது இவை இரண்டு பாட்டானும் கூறப்பட்டது.)

System resources and converting them to Tamil speech:

Besides the above online resources, it is also quite possible to interact with the system resources such as time and date and have this device record in text as well as speak it in Tamil as shown below:

Input: இப்ப □ணி என்ன
Output: இப்பொழுது அ□ரிக்கக் கிழக்கு நேரம் □ாலை எட்டு □ணி ஒன்பது நிமிடம் ஐம்பத்து ஒன்று வினாடி

Input: இன்னைக்கு என்ன தேதி
Output: இன்றைக்கு இரண்டாயிரத்து இருபத்து மூன்றுஆம் வருடம் பத்துஆம் □ாதம் பதின் ஒன்றுஆம் நாள்

In order to convert Tamil numbers to corresponding Tamil text the resource from http://robot.tamilnlp.com/py/convert_tamil_number.py is extensively used as part of this device.

Conclusion:

With this initial analysis with a robot on the speech to text and text to speech resources that are available online, an attempt is made to test the performances of the robot (cf. Renganathan 2022) in a multiple number of ways. What is yet to be attempted, but in the process of being developed, is the process of analyzing output text from speech to text resources in a meaningful way to build some of the NLP tasks such as question-answering systems, man-machine interactions along the line of natural conversations and so on. Attempting to decipher the correct interpretation of commands involving ambiguous words would be a challenging task. As already mentioned, such tasks can be accomplished only when the training is made with extensive database containing all possible bi-directional predictable expressions. Capturing the nuances of expressions involving homonymous words, semantically extended phrases etc., are to be accounted for in a precise manner possible so further advances can be made. Such projects would mainly explore the intersection between the theoretical knowledge of linguistics and the linguistic performances related to the recent advances of AI particularly in the context of building LLM and development of vector databases. Obviously, as one can see that the linguistic performances of AI models such as Bard, Chat-GPT have made enormous successes mostly without the application of much of the knowledge from theoretical linguistics, but the outcome of these models, as has been cited in this work, requires proper application of linguistic theories further so a desired and most plausible outcome can be arrived at.

References:

Noir, Nicole 2020. A dummy's guide to Bert. (https://medium.com/swlh/bert-139acce0592d).
Renganathan, Vasu. 2021. Paper presented at the International Conference on Tamil Computing, TIC2021. "என் பேரு தமிழு (en̲ pēru tamiḻu): A Speech Recognition Robot for Tamil" - http://uttamam.org/papers/21_32.pdf - (http://robot.tamilnlp.com).
Renganathan, Vasu 2016. Computational Approaches to Tamil Linguistics. Cre-A, Chennai.

The Impact of Large Language Models (LLMs) on the World

Dr. Uthayasanker Thayasivam

Univ of Moratuwa, Moratuwa, Sri Lanka

Large Language Models are powerful artificial intelligence models that have seen significant advancements in recent years. These models are trained on vast amounts of textual data and have the ability to understand and generate coherent responses to natural language queries. These models are designed to understand and generate human language, surpassing human-level performance in various language understanding tasks. These models, built on deep learning techniques and trained on large-scale textual datasets, have the ability to understand and generate coherent and contextually relevant responses to linguistic queries.

Large Language Models have had a transformative impact on various fields, revolutionizing the way tasks are planned and completed. In the field of natural language processing and chatbots, LLMs have enhanced the capabilities of chatbots, making them more human-like in their customer interactions. By processing natural language input and generating words based on the data seen, LLMs have made chatbots more efficient and responsive. Large Language Models have also been used in the field of robotics for task planning(Gao et al., 2022).Task planning with Large Language Models has been shown to assist robots in learning novel activities and completing complex tasks. Furthermore, LLMs have proven to be effective in providing high-level semantic knowledge about the physical world and common human activities.

In the future, LLMs have the potential to reshape various fields. In the field of healthcare, LLMs can be utilized to improve medical diagnosis and treatment recommendations. (The promise of large language models in health care - The Lancet, n.d)By analyzing large amounts of medical literature and patient data, LLMs can provide valuable insights and assistance to healthcare professionals. In the field of education, LLMs can be used as powerful tools for language learning and personalized tutoring. By understanding and generating natural language, LLMs can create interactive learning experiences tailored to individual students' needs. In the field of journalism, LLMs can assist in generating news articles and analyzing vast amounts of information to uncover meaningful insights. These advancements in LLM technology have the potential to revolutionize the way news is created and consumed.

While Large Language Models offer significant potential in various fields, their implementation also poses several challenges. One of the main challenges is the issue of hallucinations. Hallucination refers to a problem where LLMs generate responses that are not based on factual information but rather on patterns learned from the extensive training data. This can lead to the

generation of false or misleading information, which can have serious consequences in fields such as journalism and healthcare. Additionally, the sheer size and computational requirements of LLMs pose challenges for their implementation. Training and deploying Large Language Models require substantial computational resources and infrastructure, making them less accessible to smaller organizations or research institutions with limited resources.

Another challenge is the issue of bias in LLMs. Bias in LLMs refers to the tendency of these models to reflect and perpetuate biases present in the training data. In the case of language models trained on text from the internet, biases in the data can be amplified and reinforced in their outputs. This can lead to biased and discriminatory language generation, which can have negative consequences in various domains, including healthcare, law, and social discourse.

LLMs hold great potential for advancing the field of linguistics. These models can be used to analyze and understand linguistic patterns, including syntax, semantics, and discourse structure. By training LLMs on extensive linguistic datasets, researchers can gain insights into language usage and evolution. For example, LLMs can be used to study language acquisition and development, as well as how different languages vary in their structure and usage. Furthermore, LLMs can assist in language translation and interpretation tasks. They can be trained on bilingual or multilingual datasets to improve the accuracy and efficiency of translation systems.

One of the notable impacts of Large Language Models is in the realm of low-resource languages. These languages, which have limited textual data available for training language models, often face challenges in developing accurate language processing tools. However, LLMs have the potential to bridge this gap by leveraging the vast amount of data available in major languages and transferring knowledge to low-resource languages

Large Language Models have the potential to greatly impact local languages and address some of the challenges faced by low-resource languages. By training LLMs on larger languages, such as English or Spanish, and then transferring the knowledge to local languages, these models can assist in developing accurate language processing tools for local languages. This can have several implications: increasing accessibility and inclusivity, preserving linguistic diversity, and promoting cultural heritage. Increasing Accessibility and Inclusivity: LLMs can help improve accessibility and inclusivity by providing language processing tools for local languages.

The utilization of Large Language Models raises important ethical considerations that must be taken into account. Firstly, there is a concern regarding biased data and biases within the models themselves. Language models are trained on massive quantities of text data, which can reflect biases present in society. These biases can be perpetuated and amplified by the LLMs, leading to potential discrimination or unfairness in their outputs. Additionally, there is a need to consider the privacy and security implications of LLMs.

The data used to train Large Language Models often consists of user-generated content, which may include personal or sensitive information. The impact of LLMs on different fields is transformative. In the field of healthcare, LLMs can assist in clinical decision-making by providing recommendations or generating patient reports. However, the use of LLMs in healthcare raises concerns about patient privacy and data security.

Furthermore, LLMs raise concerns related to accountability and transparency.

It is important to understand how these models make decisions and generate responses, especially in critical domains such as healthcare. Without proper transparency, it can be challenging to identify and address potential biases or errors in the outputs of LLMs. These ethical issues surrounding the use of LLMs highlight the need for careful consideration and robust guidelines.

One possible use case for LLMs in linguistics is the study and analysis of Tamil language. Tamil is a rich and ancient language with a long literary tradition. However, due to the complexity of Tamil grammar and the lack of comprehensive linguistic resources, studying and understanding the language poses challenges. LLMs can be leveraged to develop language models specifically for Tamil, allowing researchers to analyze and generate text in the language.

References:

Gao, X., Gao, Q., Gong, R., Lin, K., Govind, T., & Sukhatme, G S. (2022, October 1). DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. https://scite.ai/reports/10.1109/lra.2022.3193254

The promise of large language models in health care - The Lancet. (n.d). https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(23)00216-7/fulltext

Building Tamil AI - Open challenges and the role of the community

Abinaya Mahendiran
abinaya@nunnarilabs.com
CTO, Nunnari Labs

INTRODUCTION

The recent advancements in the Artificial Intelligence (AI) industry, specifically in the Natural Language Processing (NLP) domain, primarily focus on English and other languages that use Latin script due to the abundance of data, active research community and evaluation benchmarks. But there are nearly 7,168 modern living languages around the world as per the 26th edition of Ethnologue. Tamil is one of those morphologically rich classical languages of the world continuing to withstand the test of time. In spite of the availability of rich literary works, exclusive NLP models have never been built for Tamil. The objective of this paper is to address how the Tamil community can collaborate and build exclusive Tamil AI overcoming the challenges and technical difficulties.

OPEN CHALLENGES

There has been a growing awareness of the need for diversity in language representation in NLP. Researchers are increasingly working on extending NLP models to support a broader range of languages and scripts. There are efforts to develop multilingual models that can understand and generate text in multiple languages. The following are some of the challenges involved in building Tamil AI,

i. Data Curation: The existing SoTA multi-lingual models and Large Language Models (LLM) use different datasets that contain Tamil subset collected from Wikipedia, websites, books and online forums with little to no human validation. Tamil language has vast amounts of literature, some of which is digitised through open-source projects like Project Madurai. This digitised data should be leveraged to create task-specific datasets after being carefully curated and validated by linguists, and NLP experts who are well-versed in Tamil.

ii. Performing Fundamental Research: Tamil is a morphologically rich language which is agglutinative in nature. Most of the SoTA models' tokenizers and encoding schemes do not work well for Tamil because of the grammatical differences between English and Tamil. Though existing model architectures can be used as is for any language, fundamental components like syntactic trees, shallow and deep parsers, tokenizers, stemmers, lemmatizers and special encoding schemes to create embeddings need to be created specific to Tamil. Fundamental research has to be carried out to build better exclusive models for Tamil.

iii. Fostering Open Source Projects: Projects like Bhashini, and AI4Bbharat that aim at curating high-quality datasets for Indic languages and building SoTA Indic base models should be given utmost importance. Fostering such projects will help us build exclusive Indic language models, including Tamil models. Aya by Cohere for AI is another open science initiative aimed at curating high quality datasets for 101 underrepresented languages of the world and building a better multilingual model.

THE ROLE OF COMMUNITY

Efforts should be undertaken to bring together the community of researchers, linguists, NLP experts, volunteers and language enthusiasts. The community can be supported through grants and funds to cover the infrastructure costs involved in developing new models and be incentivized based on their needs. Incentives can be monetary, authorship, digital certificates, swags or anything that could be of value to the different groups.

CONCLUSION

By addressing the different challenges, the community will be able to contribute better to the Tamil ecosystem and help the language co-exist and evolve with technological advancements.

REFERENCES

1. Ethnologue: https://www.ethnologue.com/
2. Project Madurai: https://www.projectmadurai.org/
3. Bhashini: https://bhashini.gov.in/
4. AI4Bharat: https://ai4bharat.iitm.ac.in/
5. Aya by Cohere for AI: https://aya.for.ai/

Tholkaappiyam: The Scientific Record of The Tamil Linguistic and Culture

Selvajothi Ramalingam
Faculty of Languages and Linguistics
Universiti Malaya, Kuala Lumpur Malaysia

Tholkaappiyam, an ancient Tamil grammar work penned by Tholkaappiyar around 2700 years ago, is a monumental literary and linguistic treasure that continues to hold immense significance for the Tamil language and its speakers around the world. While it is often classified as a grammar book, its relevance extends well beyond the confines of mere linguistic analysis. This abstract aims to shed light on the unique attributes that categorize Tholkaappiyam as a scientific record of linguistics and culture, providing valuable insights into the Tamil language's structure and evolution.

This study employed a library research methodology, with an emphasis on a meticulous examination of the text itself. Data was exclusively sourced from the verses of Tholkaappiyam, referred to as *NuRpaa* research findings presented herein conclusively establish Tholkaappiyam as a document of scientific record, transcending its role as a mere grammar treatise.At its core, Tholkaappiyam systematically documents the phonology, morphology and poetics of the Tamil language and also *thiNaimarabu*. It is essential to recognize that Tholkaappiyam's contents were not arbitrarily constructed but were derived from empirical observation and meticulous analysis. This empirical foundation renders it a comprehensive and authoritative source on Tamil grammar and culture.

Tholkaappiyam exhibits all the key attributes associated with a significant research work or empirical study in the realms of linguistics, language, and culture. Firstly, it offers a clear problem statement, namely, the need to understand and codify the Tamil language. Secondly, it delineates well-defined objectives, chiefly the organization and preservation of the Tamil language's rules and structures. Thirdly, it rests upon appropriate theoretical foundations, drawing from the rich linguistic tradition and culture of the Tamil people. Fourthly, it embodies a methodical approach to data collection, encapsulating the principles, syntax, and usage of Tamil. Lastly, it presents a rigorous data analysis, contributing to a systematic understanding of the language's intricacies and its profound cultural implications.

In conclusion, Tholkaappiyam stands as a remarkable scientific record, not just of the Tamil language, but also of the culture and heritage of the Tamil-speaking people. Its enduring relevance and value extend far beyond the historical and linguistic realms, as it encapsulates the very essence of a people, their unique expression, and their intellectual achievements. Tholkaappiyam's comprehensive documentation of Tamil grammar and literature, underpinned by empirical observation and analysis, solidifies its place as an invaluable cultural and linguistic relic. It is a testament to the timeless importance of preserving and understanding the heritage of one's language, one's culture, and the collective wisdom of generations past, making it an enduring source of inspiration for the Tamil-speaking world and a testament to the universality of linguistic and cultural exploration.

Keywords: scientific record,Tamil language, Tamil linguistic and Tamil culture

Building transformer-based models for natural language processing applications

Dr S. K. Lavanya (Anna University MIT Campus, Chennai)

Natural Language Processing, A.I, M.I

In my presentation, I will give an overview of how large language models (LLMs) based on Transformers can be used to edit, examine, and produce text-based data as well as how programmers can use these LLMs to build powerful NLP systems that allow for easy and natural human-computer interactions in chatbots, AI speech agents, and other applications. We will look at how BERT, a transformer-based LLM, has revolutionized NLP by producing results for question answering, entity recognition, intent recognition, sentiment analysis, and other tasks . We will go over how to use these models for different NLP tasks, like text classification, named entity recognition (NER), and answering questions.

Generative AI in the context of Tamil Language

Dr. Subalalitha C N, Associate Professor, SRM University

Natural Language Processing, Machine Learning, Discourse Analysis and Computational Linguistics

---

Information Extraction from Tamil Medicinal Documents

Prof D. Thenmozhi,
Department of Computer Science at SSN College of Engineering, Chennai, India
Member of the Machine Learning Research Group of SSN

Dr. Uthayasanker Thayasivam
PhD (U. Georgia), BSc Eng. (Hons) (Moratuwa)

Hands-on experience and knowledge in Data Science and Big Data projects related to text mining, and automatic extraction of ontologies. Experience in mentoring students & leading research scientists towards applying data science in decision making. Exposure, experience, and contacts during long industrial (with Ask search engine) and academic (PhD) career.

துல்லியமான விவசாயம் மற்றும் கால்நடை மேலாண்மைக்கு செயற்கை நுண்ணறிவும், ஆழக்கற்றலின் பயனு

செல்வ முரளி

ஆய்வுச்சுருக்கம்

இன்றைய விவசாயத்திற்கு நிறைய சிக்கல்கள் இருந்தாலும் குறைந்தது 3 வித சிக்கல்கள் மிக முக்கிய□ானவைகளாக இருக்கிறது.

உலக அளவில் தொழில் செய்பவர்களில் (சிறு தொழில் முதற்கொண்டு பெரிய தொழில் வரை) 80% பேர் பருவநிலை □ாற்றத்தால் பாதிப்புக்கு ஆளாகக்கூடும் என்று எச்சரிக்கப்படுகிறது. இந்தியாவில் 94% பேர் பருவநிலையை நம்பித்தான் வியாபாரம் செய்கின்றனர். இவ்வளவு அதிக□ானோர் பருவநிலையை நம்பியிருக்கும்போது, ந□க்கு என்ன தேவைகள் இருக்கின்றன, இப்போது இருக்கும் சேவைகள் நம் தேவைகளைப் பூர்த்திசெய்கின்றனவா என்று பார்த்தால் நிச்சய□ாக இல்லை என்பதே யதார்த்தம்.

ஒரு □ாவட்டத்தின் எந்தெந்தப் பகுதிகளில் □ழை பெய்யும் என்று துல்லிய□ாகக் கணித்துச் சொல்ல முடியவில்லை. பொதுவாக, □ாவட்டந்தோறும் வானிலை முன்னறிவிப்பு செய்யப்படுகிறது. ஆனால், அன்று □ாவட்டம் முழுவதும் □ழை பெய்கிறதா என்றால் இல்லை. ஒரு கிரா□த்தில் □ழை பெய்தால், அடுத்த கிரா□த்தில் பொழிவதில்லை, அதனால் என்ன பாதிப்பு நேர்ந்துவிடப்போகிறது? சில உதாரணங்களைப் பார்ப்போம்.

அறுவடைக்கு முன்னர், அதிக □ழை பெய்யும் அல்லது அதிக வெப்பம் ஏற்படும் என்று விவசாயிகளுக்குத் துல்லிய□ாக எச்சரிக்கை கொடுக்கப்படுவதில்லை. பயிர் விளையும்போது அதிக □ழை பெய்தால் பயிர்களில் அழுகல் நோய் வரலாம், அவை தொற்றுக்குள்ளாகலாம் அல்லது சாய்ந்தேவிடலாம். அறுவடைக்குப் பின்பு நெற்பயிர்களைக் காய வைக்கும்போதோ வித்து வகைகளைக் காய வைக்கும்போதோ திடீரென □ழை பெய்தால் விவசாயிகள் செய்வதறியாது திகைத்துப் போய்விடுகின்றனர்.

Tamil Text Generation using ChatGPT-3 Models

Dr. R. Ponnusamy
Dept of Computer Science and Engineering
Chennai Institute of Technology
Email: ponnusamyr@citchennai.net

Abstract: Tamil Text Generation using GPT-3 Models is a fascinating application of advanced

natural language processing technology. GPT-3, a generative pre-trained transformer, has been trained on extensive Tamil language data, enabling it to generate coherent and contextually relevant Tamil text. The process begins with a user providing a prompt or input text in Tamil. GPT-3 leverages its understanding of the language's grammar, vocabulary, and contextual cues to generate text that logically extends from the input. It's capable of generating anything from short sentences to lengthy paragraphs, adapting its output to the provided context. What sets GPT-3 apart is its ability to produce human-like and creative text. It can answer questions, create engaging stories, compose articles, and even write code in Tamil, making it a versatile tool for content generation across various domains.
Moreover, GPT-3 can be fine-tuned for specific tasks or industries, enhancing its performance in specialized applications. This adaptability has led to its use in chatbots, content creation, automated customer support, and more, where it can save time and effort in generating high-quality text. However, it's crucial to exercise caution when using GPT-3 for text generation. Ethical concerns, such as bias in generated content, the potential for spreading misinformation, and the need for human oversight, must be addressed. Tamil Text Generation with GPT-3 models is a powerful tool that harnesses the capabilities of pre-trained language models to generate coherent and contextually relevant Tamil text for a wide range of applications, revolutionizing content creation and language-related tasks in the Tamil language. In this paper, an attempt is made to understand the basic design of the ChatGPT-3 model in a detailed manner.
Keywords: Tamil Text Generation, GPT-3 Models, chatbot, trained language models, natural language processing.

1. Introduction

Natural Language Processing is a challenging task in the world. In a human brain, it is natural to store various real-world objects in different forms, that is, in the form of text, audio and video. Also, internally, a human can visualize the task and things that they are seeing in the real world. Able to get attention while speaking with others and answer the questions as well. In the real world, it is a significant challenge to create such systems. There are several attempts have been made to develop such strategies in the computational world. One such Chatbot is ChatGPT (Chat Generative Pre-trained Transformer), which is an interactive text-processing system which is capable of generating human-like text responses, engaging in conversations, and providing assistance with various topics and tasks based on the input it receives. It also performs functions like Conversational Interaction, Answering Questions, Language Translation, Content Creation, Tutoring and Learning, Programming Help, idea Generation, etc.
In the abstract, one can easily explain the task, but the creation of such a system and understanding the design is a complex task. In this paper, it explains the procedure and nature of the working of Chat GPT – 3 models in a detailed manner. Section 2 gives a literature survey about ChatGPT usage and model—Section 3 Explains the working nature of the operating model of ChatGPT. Section 4 explains the attention model of Transformers modelling, Section 5 gives the empirical modelling of the self-attention system, and Section 6 concludes the paper.

2. Literature survey

The reference points to a book titled "Natural Language Processing with Transformers", authored by Tunstall, Werra, and Wolf, published by O'Reilly Media in 2022 [1]. This book delves into the field of natural language processing, specifically focusing on transformer models. It likely covers the theory, techniques, and practical applications of using transformers in processing and

understanding human language. Readers can expect comprehensive insights into the cutting-edge methods of natural language processing, with a particular emphasis on transformer architectures, making it a valuable resource for developers, researchers, and NLP enthusiasts.

Nazir and Wang, in 2023[2], present a comprehensive survey on ChatGPT, focusing on its advancements, applications, prospects, and challenges. Published in Meta-Radiology, the study explores the evolving landscape of ChatGPT technology. It discusses recent advances in the field, detailing the diverse applications of ChatGPT in various sectors. The paper critically assesses the prospects of ChatGPT, shedding light on its potential developments. Additionally, the research addresses the challenges faced by ChatGPT, providing insights into the hurdles that need to be overcome for further progress. The paper serves as a valuable resource for scholars, researchers, and professionals interested in understanding the current state and future potential of ChatGPT technology.

In their 2023 paper, Raj, Singh, Kumar, and Verma [3] investigate the potential advantages and applications of ChatGPT for enhancing the efficiency and effectiveness of business operations. Published in Bench Council Transactions on Benchmarks, Standards, and Evaluations, the research delves into the role of ChatGPT as a tool in business contexts. The study explores various use cases, demonstrating how ChatGPT can be leveraged to streamline and optimize a range of business processes. This work is valuable for business professionals seeking to understand the practical implications of ChatGPT technology in their operations.

In their 2023 preprint, Rahman and Watanabe [4] explore the implications of using ChatGPT in education and research contexts. The study delves into the opportunities presented by ChatGPT, highlighting its potential benefits in enhancing educational experiences and advancing research methodologies. The authors also address the associated threats, discussing ethical concerns and possible limitations of integrating ChatGPT in educational settings. The paper offers strategic insights, proposing approaches and guidelines to maximize the advantages of ChatGPT while mitigating risks effectively. This research is invaluable for educators, researchers, and policymakers aiming to harness AI technologies for educational and research purposes.

In their paper presented at the 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Bhardwaz and Kumar [5] conducted a comprehensive comparative analysis of three prominent chatbot technologies: ChatGPT, Google BARD, and Microsoft Bing. The study delves into the intricacies of these technologies, evaluating their functionalities, performance, and capabilities. Through this analysis, the authors aim to provide valuable insights into the strengths and weaknesses of each chatbot system. This research is essential for professionals and researchers in the field of artificial intelligence, offering a detailed understanding of the comparative landscape of these chatbot technologies.

In their 2023 study published in the Mesopotamian Journal of Computer Science, Sakirin and Ben [6] they have investigated user preferences regarding ChatGPT-powered conversational interfaces in comparison to traditional methods. The research focuses on understanding how users perceive and interact with ChatGPT-powered systems compared to conventional methods of communication. By exploring user preferences, the study sheds light on the acceptance and usability of ChatGPT in real-world conversational scenarios. The paper likely discusses findings related to user satisfaction, ease of use, and effectiveness of ChatGPT-powered interfaces when contrasted with traditional methods. This research contributes valuable insights to the field of human-computer interaction, providing a nuanced understanding of user preferences in the context of advanced conversational interfaces.

The reference highlights the research conducted by A. Baki Kocaballi [7] at the School of

Computer Science, University of Technology Sydney. The study focuses on Conversational AI-powered design, specifically utilizing ChatGPT in various roles: as a designer, user, and product. This research likely explores the innovative applications of ChatGPT in the realm of design processes, demonstrating its versatility and potential impact on user experience and product development. The work sheds light on the intersection of artificial intelligence and design, offering valuable insights into the evolving landscape of AI-driven creative processes.
The referenced book, "GPT-3: Building Innovative NLP Products Using Large Language Models," authored by Sandra Kublik and Shubham Saboo in 2022[8], explores the practical applications of GPT-3, a large language model developed by OpenAI. The book likely delves into the techniques and methodologies for building innovative Natural Language Processing (NLP) products using GPT-3. Readers can expect insights into leveraging this advanced language model for creative and impactful NLP applications, providing a valuable resource for developers, researchers, and professionals interested in harnessing the power of large language models for innovative projects.
The reference discusses a guide titled "Building Transformer Models with Attention," authored by Jason Brownlee, Stefania Cristina, and Mehreen Saeed in 2022 [9] for Machine Learning Mastery. The guide demonstrates how to create a Neural Machine Translator from scratch using the Transformer architecture in Keras. It provides practical insights and implementation techniques for developing sophisticated neural network models, explicitly focusing on attention mechanisms. This resource is valuable for developers and machine learning enthusiasts seeking a hands-on understanding of building advanced models, showcasing real-world applications of Transformer architectures in the field of machine translation. In the above literature survey, an analysis is made to study the different aspects of the ChatGPT System.

3. Operating Model of ChatGPT

OpenAI is a research organisation founded in 2015 with the goal of promoting and developing friendly AI that benefits humanity. Later, in 2018, OpenAI introduced GPT (Generative Pretrained Transformer), a transformer-based language model that was trained on a large corpus of text data. ChatGPT, which stands for Chat Generative Pre-trained Transformer, is a Large language model-based chatbot developed by OpenAI and launched on November 30, 2022.
ChatGPT is built upon either GPT-3.5 or GPT-4—members of OpenAI's proprietary series of generative pre-trained transformer (GPT) models. Initial approaches focused on rule-based systems and hand-crafted linguistic models and later used deep learning methods have enabled the development of more sophisticated and effective language models. Based on the transformer architecture developed by Google—it is fine-tuned for conversational applications using a combination of supervised and reinforcement learning techniques.
Foundation models are trained with a wide variety of data and can transfer knowledge from one task to another. It contains hundreds of billions of hyperparameters that have been trained with hundreds of gigabytes of data. BLOOM (Big-Science Large Open-science Open-access Multilingual Language Model) is a critical foundation model created by volunteers from a community-driven machine learning (ML) platform called Hugging Face. The BLOOM model, which included 176 billion parameters and was trained for 11 weeks, is now available to the public and can be accessed through the Hugging Face website. The Centre for Research on Foundation Models (CRFM) is a new interdisciplinary initiative born out of the Stanford Institute for Human-Cantered Artificial Intelligence (HAI) that aims to make fundamental advances in the study, development, and deployment of foundation models. Foundation models (e.g., BERT, GPT-3, CLIP, Codex) are models trained on broad data at scale such that they can be adapted to a wide

range of downstream tasks.
The working model of the ChatGPT is given in the following figure 1. It is a three-step process.

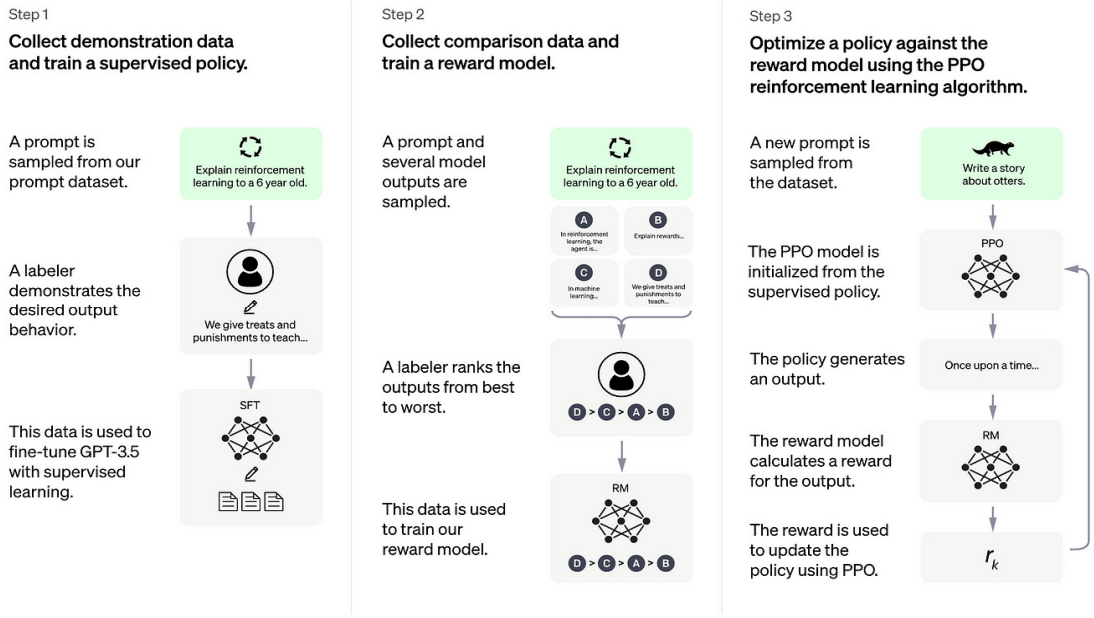


Figure 1 ChatGPT Working Model

Step 1 Pre-Training

i. Architecture: ChatGPT is based on a transformer architecture. Transformers are deep-learning models designed to handle sequential data efficiently.

ii. Pre-training Corpus: Initially, the model is pre-trained on a massive corpus of text data from the internet. It learns to predict the next word in a sentence, given all the previous terms. This process helps the model learn grammar, facts, reasoning abilities, and some biases present in the training data.

iii. Unsupervised Learning: During pre-training, the model learns to represent language patterns in a way that allows it to generate coherent and contextually relevant responses to given prompts.

Step 2 Fine-tuning

i. Custom Datasets: OpenAI fine-tunes the model using custom datasets created by OpenAI. These datasets include demonstrations of correct behaviour and comparisons to rank different responses.

ii. Human Feedback: Human reviewers assess and rate model outputs for a range of example inputs. OpenAI uses this feedback to create a reward model, which is used to fine-tune the model further.

iii. Iterative Process: Fine-tuning is an iterative process where the model is repeatedly adjusted and evaluated based on human feedback until it performs well according to OpenAI's defined criteria.

Step 3 Reinforcement Learning from Human Feedback

Fine-tuning ChatGPT with RLHF consisted of three distinct steps:

i. Supervised fine-tuning (SFT) – A pre-trained language model is fine-tuned on a relatively small amount of demonstration data curated by labellers to learn a supervised policy (the SFT model) that generates outputs from a selected list of prompts. It represents the baseline model.

ii. "Mimic human preferences" – Labellers are asked to vote on a relatively large number of the SFT model outputs, this way creating a new dataset consisting of comparison data. A new model is trained on this dataset. It is referred to as the reward model (RM).

iii. Proximal Policy Optimization (PPO) – The reward model is used to fine-tune further and improve the SFT model. The outcome of this step is the so-called policy model.

4. Transformer Model Self-Attention

The ChatGPT uses the transformer architecture to generate the text. It is a neural network architecture. The main crux of the transformer architecture is the multi-head attention model, which compares to the self-attention model. The architecture is shown in the following figure 2.

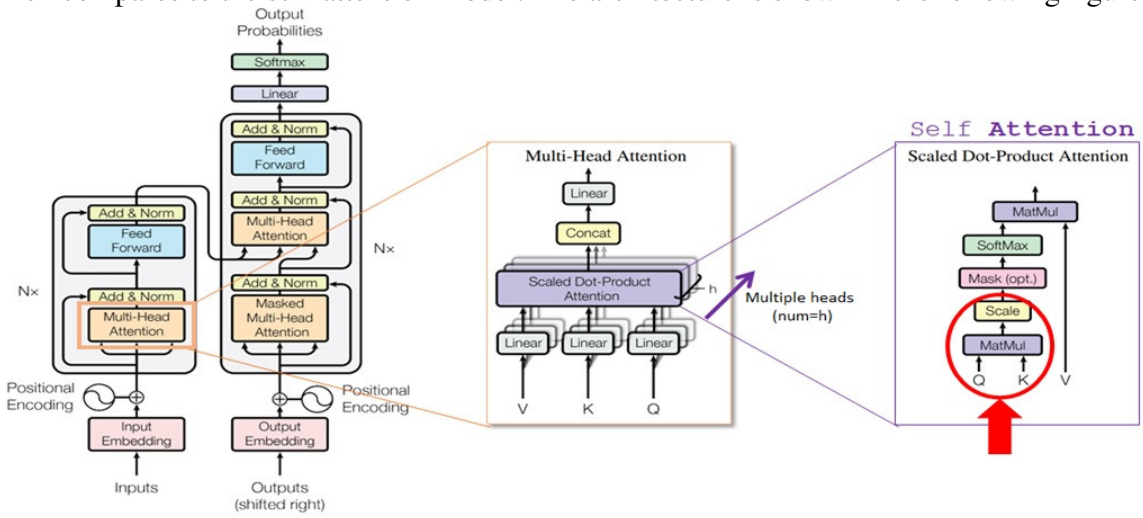


Figure 2 Transformers architecture – Multi-Head Attention and Self–Attention Model

5. Empirical Modelling of Self-Attention Modelling

In a human brain, when you read a sentence, a process happens for each word in the sentence as your eyes progress through the sentence. For example, the sentence "*ராஜா ஆப்பிரிக்கா வந்தார்*". When your eyes see *ராஜா*, your brain looks for the most related word in the rest of the sentence to understand what *ராஜா* is about (query). Your brain focuses or attends to the word *வந்தார்* (key).

This process is implemented through scaled dot-product attention; the input sequence was transformed using three matrices representing the query, key, and value.

The first MatMul implements an inquiry system or question-answer system that imitates this brain function, using Vector Similarity Calculation.

Think of the MatMul as an inquiry system that processes the inquiry: "For the word q that your eyes see in the given sentence, what is the most related word k in the sentence to understand what q is about?" The inquiry system provides the answer as the probability.

Table 1 Query – Key Probability Matrix

q	k	probability
ராஜா	வந்தார்	0.94
வந்தார்	ஆப்பிரிக்கா	0.86
ஆப்பிரிக்கா	வந்தார்	0.76

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The next word to be generated is computed in a self-attention system is computed through the above model. It is shown in the following figure 3.

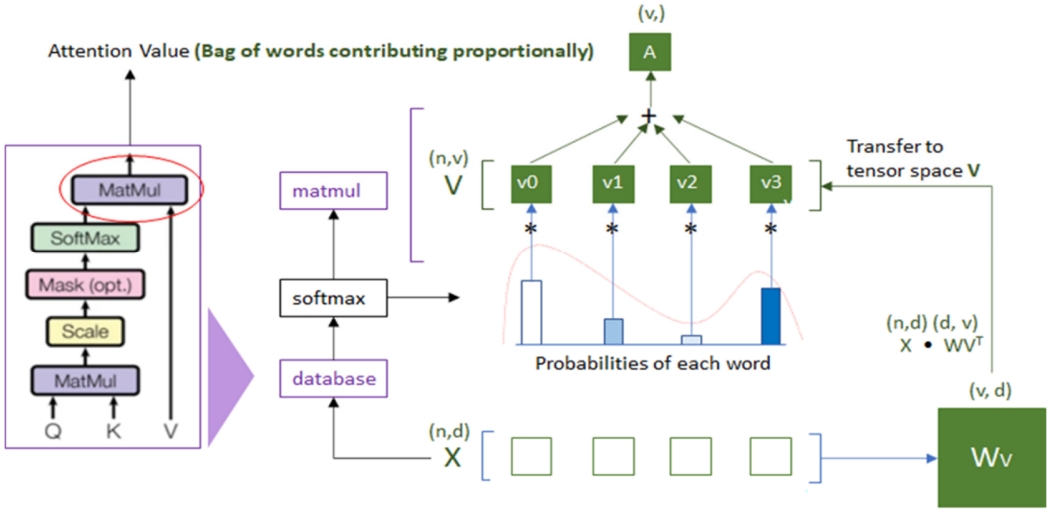


Figure 3 Attention Value Computation

6. Conclusion

In this paper it gives the complete working nature of the ChatGPT model and the process of bootstrapping the ChatGPT model for different languages, especially in Tamil. Understanding the Complete ChatGPT design requires learning about various components and learning methods. The model uses supervised, unsupervised and reinforcement learning methods. It also explains the transformer. One of the essential components of Transformer learning is attention gaining or attention computation. The learning-based self-attention model is presented, and its working nature is presented with an example.

7. References:

1. Tunstall, L., Werra, L. V., & Wolf, T. (2022). Natural language processing with transformers. O'Reilly Media.

2. Nazir, A., & Wang, Z. (2023). A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. Meta-Radiology, 100022. https://doi.org/10.1016/j.metrad.2023.100022.
3. Raj, R., Singh, A., Kumar, V., & Verma, P. (2023). Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(3), 100140. https://doi.org/10.1016/j.tbench.2023.100140
4. Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. https://doi.org/10.20944/preprints202303.0473.v1.
5. Bhardwaz, S., & Kumar, J. (2023). An extensive comparative analysis of chatbot technologies - ChatGPT, Google BARD and Microsoft Bing. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). https://doi.org/10.1109/icaaic56838.2023.10140214
6. Sakirin, T., & Ben Said, R. (2023). User preferences for chatgpt-powered conversational interfaces versus traditional methods. Mesopotamian Journal of Computer Science, 24-31. https://doi.org/10.58496/mjcsc/2023/006
7. A. Baki Kocaballi, Conversational AI-Powered Design: ChatGPT as Designer, User, and Product,
School of Computer Science, University of Technology Sydney, baki.kocaballi@uts.edu.au
8. GPT-3: Building Innovative NLP Products Using Large Language Models, ISBN:9355422024, Sandra Kublik, Shubham Saboo, 2022.
9. Building Transformer Models with Attention: Implementing a Neural Machine Translator from Scratch in Keras. Jason Brownlee, Stefania Cristina, Mehreen Saeed, Machine Learning Mastery, 2022.

International Conference on Tamil Computing, KCT, Oct 13-14, 2023

List of Abstracts received for the young researchers paper presentations

1. **Porul Thaedal - A Tensor-Based Semantic Representation Model for Enhancing Tamil Language Understanding.**
   Muthu Vignesh, Yugeshwaran, Deeptharun & Kannan C
   KIT- Kalaingnar karunanidhi Institute of Technology

2. **Tamizhi Inscriptions Lexicon for Machine Translation**
   Monisha Munivel[1*], V S Felix Enigo[2], Suresh Balaji K[3]
   SSN College of Engineering

3. **Question answer retrieval for thirukkural**
   Roshan B & Mohamed Saffi M
   Thiagarajar college of engineering

4. **Linguistic Translator**
   Vidhya Kanagaraj
   KG College of Arts and Science

5. **Legal Assistant Through AI Chatbot In Tamil For Cyber Crimes Against Women**
   Dr. S.K Lavanya Assistant Professor & *Shriya S, Jayasimman J*
   Madras Institute of Technology

6. AI-Powered YouTube Transcript Summarization with Transformers models

V.Shanmugapriya[1] V.Srividhya[2]
Avinashilingam Institute for Home Science and Higher Education for Women

7. Code-mixed "computationally romba challengingaa irukku"

Kathiravan Pannerselvam and Saranya Rajiakodi
Central University of Tamil Nadu-Thiruvarur

8. Deep Learning for Sarcasm Identification in Tamil-English Code-mixed Data

Ramya priya S[1], Shanmitha Thirumoorthy[2], DurairajThenmozhi[1]
SSN College of Engineering

9. தொல்காப்பியக் குறுஞ்செயலி உருவாக்கம் / App Development for Tholkaappiyam

முனைவர் வினோத் அ., பூவேந்திரன் கோ., முனைவர் சத்தியராஜ் தங்கச்சாமி,
கணினித் தொழில்நுட்பவியல் துறை, ஸ்ரீ கிருஷ்ணா ஆதித்யா கலை மற்றும் அறிவியல் கல்லூரி, கோயமுத்தூர் - 641042

10. "Exploring Tamil Sentiments: Discovering 'Meipaadu' with AI in Social Media"

Dr.Balamurugan.V.T, Dhayanithi.A, Akash.S, Ramkumar.K.
Bannari Amman Institute of Technology

11. AI Based Tamil Palmleaf Manuscript Reading software

Pravin Savaridass M, Udhaya Moorthy S J, Gokul S
Bannari Amman Institute of Technology

Porul Thaedal - A Tensor-Based Semantic Representation Model for Enhancing Tamil Language Understanding.

Muthu Vignesh, Yugeshwaran, Deeptharun & Kannan C
KIT- Kalaingnar karunanidhi Institute of Technology

ABSTRACT

Language serves as a fundamental communication tool, and each linguistic entity possesses distinct grammar rules and literary traditions. Tamil, an ancient language, boasts a rich history spanning over 2000 years, characterized by unique grammatical rules and a vast literary heritage. Tamil also exhibits a remarkable richness in morphology, allowing for the creation of numerous words through the addition of morphological suffixes to a single base word. For instance, combining "மரம்" (tree) with "மா" (like) yields "மரமா" (tree-like), and "மரம்" combined with "களிலிருந்து" (from) results in "மரங்களிலிருந்து" (from the trees), among many other possibilities. However, existing language models often fail to recognize the intricate morphological variations of individual words.

Furthermore, the Tamil language exhibits an exceptionally high lexical diversity, with multiple words conveying the same meaning. For instance, the concept of "love" can be expressed through words such as "அன்பு" (anbu), "காதல்" (kaadhal), "நேசம்" (nesam), "நே" (ne), "நேயம்" (neyam), "பாசம்" (paasam), and many more. Unfortunately, conventional search engines and language models often lack the capability to recognize synonymous expressions, necessitating the need to accurately map and correlate such words.

This research introduces a novel approach to address these linguistic challenges through the development of a Tensor Model for Tamil Language Semantic Representation. Leveraging the intricate morphology and complex syntactic structures inherent to Tamil, this model represents Tamil words as multidimensional vectors, with each dimension corresponding to a unique semantic feature. By projecting Tamil words into a high-dimensional space, the model effectively captures the intricate relationships between words and their contextual meanings.

To ensure the model's robustness, it was meticulously trained on an extensive corpus of Tamil text. Subsequently, a comprehensive evaluation utilizing established semantic representation metrics demonstrates that the Tensor Model for Tamil Language Semantic Representation outperforms existing models. Notably, it offers a more precise and nuanced representation of Tamil language semantics.

This pioneering model holds significant promise for applications across the spectrum of natural language processing, machine translation, and various other fields where a comprehensive understanding of Tamil language semantics is imperative. Specifically, it can facilitate the exploration of literary works by enabling the identification of synonymous expressions across poems, songs, and other forms of Tamil literature, thereby enriching the analysis and interpretation

of this ancient and vibrant language.

Tamizhi Inscriptions Lexicon for Machine Translation

Monisha Munivel[1*], V S Felix Enigo[2], Suresh Balaji K[3]

[1] Research Scholar, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai [2] Associate Professor, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai [3] PG Student, Department of Anthroplogy, Indira Gandhi National Open University, New Delhi

[1]moni.munivel@outlook.com , [2]felixvs@ssn.edu.in, [3]ksbalaji23@gmail.com

Abstract

Machine translation is used in a variety of contexts, including business, industry, domain specific, and multi-domain. To translate from one language to another, machine translation of bilingual dictionaries is necessary. Lexicon, collection of words and the knowledge associated with their usage in a language, are found in dictionaries. To translate Tamizhi into south Indian languages, this research suggests a hybrid strategy for creating lexicons for Tamizhi inscriptions. One of the earliest inscriptions is Tamizhi, which dates to the third century BCE. One of the challenges in digitizing Tamizhi inscriptions is due to its language complexities. Tamizhi inscriptions lack spaces and dots, making it challenging to distinguish between words in the writing. A Tamizhi text is read based on the phonemes. Based on this concept, the pronunciation of Tamizhi characters is generated using a grapheme-to-phoneme (G2P) conversion approach. This is accomplished using a sequence-to-sequence (seq2seq) architecture for G2P. In this, CNN is utilized as a text-to-phoneme encoder, while Bi-LSTM is employed as a phoneme-to-text decoder. Before the morphemes are decoded to text, phonological errors are fixed after encoding in order to interpret them. A reliable machine-readable Tamizhi dictionary is created by manually validating and correcting the generated lexicons by subject-matter specialists. Any south Indian language can be translated from Tamizhi using this dictionary. So that the general population can understand the information found in Tamizhi inscriptions in their own languages, such as medical scripts, architectural notes, trade, and commercial secrets, etc.

Keywords: Machine Translation, Tamizhi Inscriptions, Lexicon, G2P Conversion, Morphemes

Question answer retrieval for thirukkural

Roshan B, Mohamed Saffi M
Thiagarajar College of Engineering

Abstract:

The Thirukkural, a classical Tamil text authored by Thiruvalluvar, comprises 1,330 couplets that offer profound insights into various aspects of life, ethics, and governance. This project presents an innovative approach to enhance the accessibility and understanding of the Thirukkural using large language model(LLM) techniques.

Our project leverages state-of-the-art large language models, specifically the BERT-based language model, to provide users for exploring the Thirukkural. Users can input their questions in Tamil, and the system will employ a large language model to analyse the input, retrieving the most relevant couplets from a curated dataset.

By bridging the gap between ancient Tamil literature and modern technology, this project aims to promote the timeless wisdom of the Thirukkural to a wider audience and facilitate a deeper comprehension of its teachings.

In the context of our l model, it was observed that BERT consistently outperformed other adapter-based approaches in terms of accuracy. Additionally, our project emphasizes that when compared to fine-tuning models, Adapter models demonstrate an advantage by necessitating the training of fewer parameters.

Text-To-Speech, Voice Recognition and Speech Corpora, Particularly in Application for Physically Challenged ("Assistive Technologies")

HemalathaK[1] (AssistantProfessor, Department of Computer Science),

Vidya K[2], Reesha G[3], Bharathkumar S[4], Shri Ram SR[5], Mohamed Hakeem S[6], Harish AG[7].

(II B.Sc Computer Science Students) ,KG College of Arts andScience, Coimbatore.

hemalatha.k@kgcas.com[1],vidyakanagaraj26@gmail.com[2]

ABSTRACT

Now a days the AI (Artificial Intelligence) Technologies is used in many fields. But in majority of medical field, many disabled and physically challenged people are expecting the help from AI field. This work probes into three essential technologies such as Text-to-Speech (TTS), Voice Recognition (VR), and Speech Corpora. It underscores their crucial contribution to the creation of assistive devices for people with physical limitations.

AI has been opening up new and simpler ways to manage our daily activities, with big potential to automate tasks that typically requires human intelligence such as a speech and voice recognition. AI can help individuals with disabilities by making a major difference in their ability to get around and take part in the activities of daily living AI-voice assisted technologies like echo, Google home, Alexa have created new means of accessibility for disabled people. As artificial intelligence took an important role in communication and interaction, the use of this technology enables individuals with disabilities to access information much easier all just by speaking to their devices.

AI can be used to develop assistive technologies that can help people with disabilities to perform tasks that would otherwise be difficult or impossible for them in this AI the people those who are not able to see can use voice recognition technology for their comfort, “By This,disabled people get benefited”,namely AI powered devices like speech recognition software and smart home devices can help people with mobility or speech impairments to communicate and control their environments.

Keywords:AI,VR,TTS

INTRODUCTION

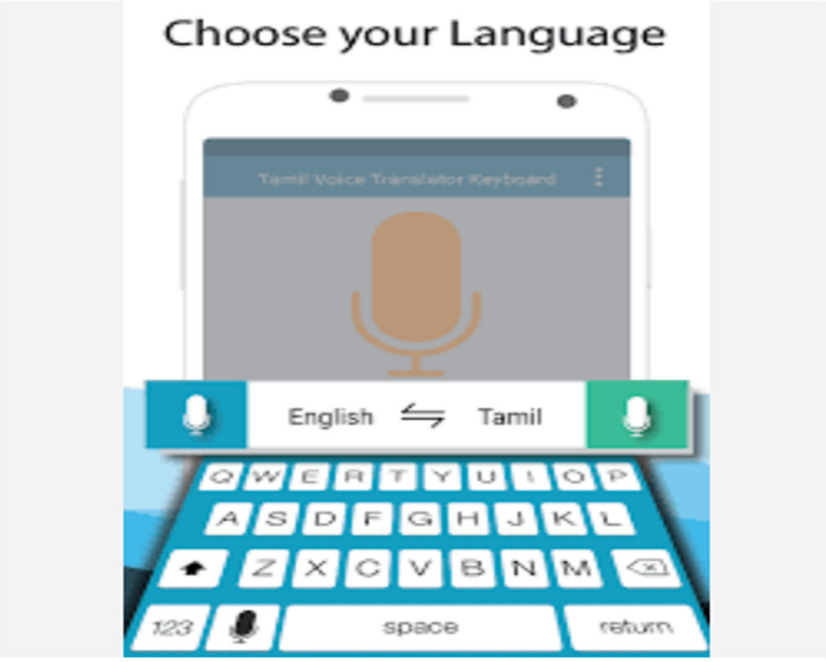
A translator is a person or tool that converts information, from one language into another, ensuring

that the content, context, and style of the original message are preserved. This process requires a deep understanding of both the source and target languages, as well as the cultures they represent, in order to convey the meaning accurately and sensitively. While human translators specialize in understanding nuances, idioms, and cultural contexts, machine translators utilize algorithms and large datasets to generate translation.

In the age of rapid technological advancements, the intersection of voice-based technologies and assistive solutions presents a paradigm shift for the physically challenged. This paper delves deep int three pivotal technologies: Text-to-Speech (TTS), Voice Recognition (VR), and Speech Corpora, and their transformative impact in crafting assistive technologies tailored for those with physical impairments.

Text-to-Speech (TTS) solutions empower users by converting written text into natural-sounding audio. This has opened doors for those who may struggle with reading due to visual impairments or specific learning disabilities, allowing for equal access to information and digital content. Voice Recognition, or speech-to-text, on the other hand, turns the spoken word into written form, paving the way for hands-free computing and aiding those who might find typing or using a touchscreen challenging. Meanwhile, the unsung hero behind these functionalities, Speech Corpora, provides the vast datasets that enable the precision and adaptability of these voice technologies, ensuring they evolve and adapt to users' needs.

For individuals with physical challenges, these technologies are not merely convenience tools. They represent newfound independence, equality, and the breaking down of barriers in a predominantly digital world. This paper aims to explore the nuances of these innovations, detailing their technical intricacies, current applications, and potential future trajectories in the realm of assistive technologies.



RELATED WORKS

Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR model is mentioned in [1]. In this model automatic speech recognition is a tool used to transform human speech into a written form. In this paper we describe an automatic speech recognition model, determined by using three pretrained models, fine-tuned from the Facebook XLSR Wav2Vec2 model, which was trained using the Common Voice Dataset. The best model for speech recognition in Tamil is determined by finding the word error rate of the data.

The work concentrates on the device, which helps as a translation system for translating sign gestures into text mentioned in [2]. "Tamil sign language translator to solve this problem." Here, gestures are translated to Tamil language to find a localized solution.

The focus of the research is to analyse real-time sign language translators that are used for language translation. Sign Language Translation Systems that were developed from 2017 to 2021 are analysed in this paper mentioned in[5]. Index Terms—Sign Language, Sign Language Recognition, Handicapped aids, Application Program Interfaces, IoT.

ARA will read out the content of the website and then using speech to text and text to speech modules along with selenium, the software can automate any website mentioned in [4]. The designed voice assistance connects with the intended applications to provide results that the user has demanded. The objective of this paper is to illustrate how voice assistants are used in everyday life and to explore whether there is potential for making them accessible for people with disabilities.

Creating a system that consists of a module that initially transforms voice input to English text and which parses the sentence, then to which Indian sign language grammar rules are applied is mentioned in [3]. This is done by eliminating stop words from the reordered sentence. Indian Sign Language (ISL) does not sustain the inflections of the word. Hence, stemming is applied to vary over the words to their root/ stem class. All words of the sentence are then checked against the labels in the dictionary containing videos representing each of the word

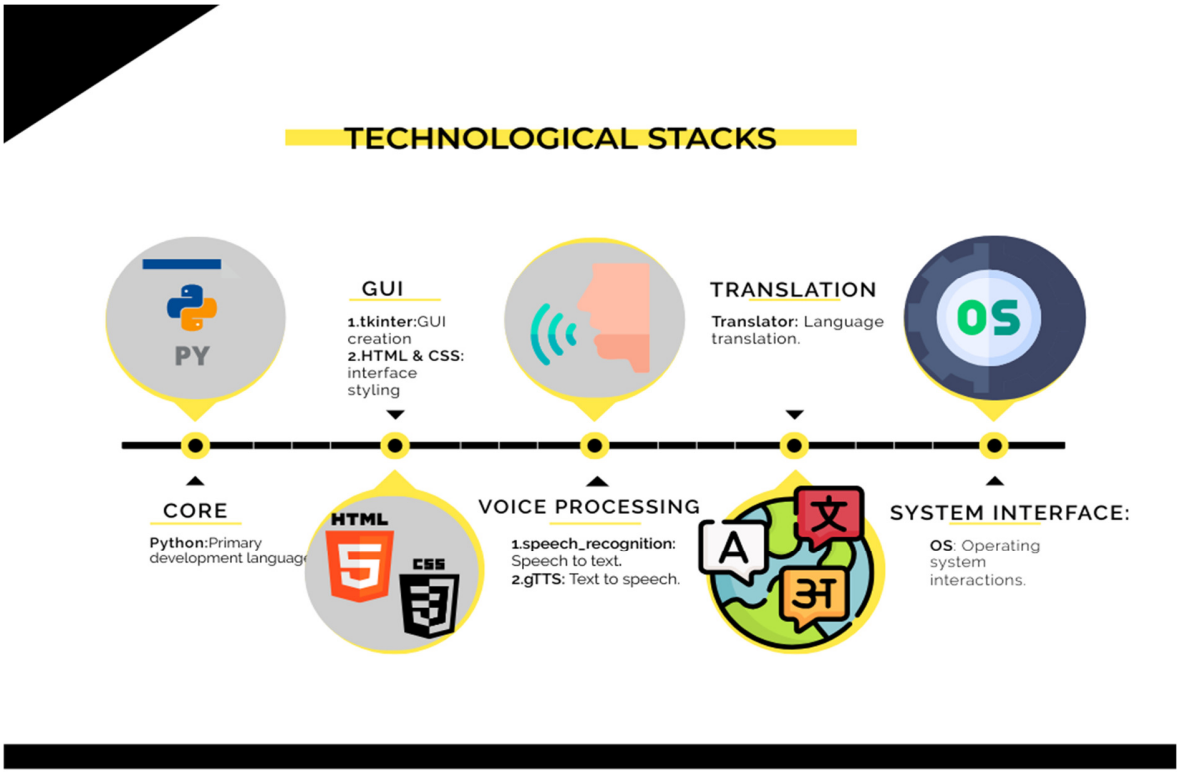
3. METHODOLOGY

The application is properly installed and all required Python modules are set up.The user has a working microphone and speakers/headphones.

Basic Flow

1. **Start Translator Application:** The user launches the translator application. The GUI,designed using '**tkinter'**and styled with HTML & CSS, displays two main sections: 'Input' and 'Output', and dropdowns for selecting languages.
2. **Select Input Language:** The user selects their native or spoken language from the first dropdown.
3. **Select Output Language:** The user selects the language they want their speech or text translated to from the second dropdown.
4. **Voice Input:** The user clicks the 'Speak' button. This action triggers the application to wait for voice input.
5. **Speech Recognition:** Once the user speaks, the application captures the voice using the '**speech recognition'** module and converts the voice input into text.

6. **Translation:** The application then takes the recognized text and translates it into the desired output language. The translation process can be performed using various translation APIs or modules available.
7. **Display Translation:** The translated text is displayed in the 'Output' section of the GUI.
8. **Text-to-Speech Conversion:** The user has an option to listen to the translated text. Upon clicking the 'Listen' button, the application uses the '**gTTS'**module to convert the translated text into speech.
9. **Play the Translated Speech:** The application plays back the translated speech using the system's default media player, facilitated by the '**os'**module.



Implementation

Input can be provided in the form of text or voice, and the output is generated in either text or voice format, making it accessible to a wide range of users.

Python Libraries

1. **tkinter**: This built-in Python library is used for creating the graphical user interface of your application.
    - Install via: Typically comes pre-installed with Python.
2. **translator**: This module provides translation capabilities. Depending on the specific library you're referring to (since there are multiple), you would need appropriate credentials, e.g., API keys.
    - Install via: **pip install translator** (or another specific package if you're referring to a different one)
3. **gTTS (Google Text-to-Speech)**: Converts the translated text into speech.
    - Install via: **pip install gTTS**
4. **os**: This built-in module allows interfacing with the underlying operating system. For this project, it might be used to play back the audio files created by gTTS.
    - Comes pre-installed with Python.
5. **speech recognition**: Recognizes speech and converts it to text.
    - Install via: **pip install Speech Recognition**

CONCLUSION

In an increasingly interconnected world, this work stands as a beacon of innovation, emphasizing inclusivity and convenience for disable peoples. By harnessing the capabilities ofAI technologies, this initiative not only benefited for disables but also bridges language barriers. For this work, future enhancements by developing a companion mobile application where it enhances the system to function effective even without a stable internet connection.

REFERENCES

[1] Dhanya Srinivasan, B. Bharathi, D. Thenmozhi, B. Senthil Kumar “SSNCSE_NLP@LT-EDI-ACL2022: Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR models” Proceedings of the Second Workshop on Language Technology for Equality, Diversity

and Inclusion, pages 317 - 320 May 27, 2022 ©2022 Association for Computational Linguistics.

[2]C.Bharathi Priya, S.P. Siddique Ibrahim, D. Yamuna Thangam, X. Francis Jency, P. Parthasarathi. "Tamil sign language translation and recognition system for deaf-mute people using image processing techniques" Proceedings of the 2023 International Conference on Software Engineering and Machine Learning.

[3] Ashmi Katariya,Vaibhav Rumale, Aishwarya Gholap, Anuprita Dhamale, Ankita Gupta. "Voice to Indian Sign Language Conversion for Hearing Impaired People" Proceedings of SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, Volume 12, Special Issue 2 (2020).

[4] Khushboo Sharma, Disha Bahal, Aman Sharma, Ankita Garg, Neeta Verma." ARA- A Voice Assistant for Disabled Personalities" Proceedings of International Journal of Engineering Applied Sciences and Technology, 2022 Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 106-109 Published Online May 2022 in IJEAST (http://www.ijeast.com).

[5] Maria Papatsimouli, Panos Sarigiannidis and George F.Fragulis. "A Survey of Advancements in Real-Time Sign Language Translators: Integration with IoT Technology" proceeding of Technologies 2023, 11, 83. https://doi.org/10.3390/ technologies11040083.

[6] S.Sudha, "Dynamic Tamil sign language recognition system," Int. J. of advanced research in management, architecture, technology and engineering, Vol.2, Issue 8, pp. 1-6, 2016.

[7] Rajat Sharma, "Communication device for differently-abled people: a prototype model," Int. conf. on data engineering and communication technology, springer, pp.565-575,2016.

[8] Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. "Findings of theshared task on hope speech detection for equality, diversity, and inclusion". In Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 61--72.

[9] R Kiran, K Nivedha, T Subha, et al. 2017. "Voice and speech recognition in tamil language". In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), pages 288--292. IEEE.

[10]A Madhavaraj and AG Ramakrishnan. 2017. "Design and development of a large vocabulary, continuous speech recognition system for tamil". In 2017 14th IEEE India Council International Conference (INDICON), pages 1--5. IEEE.

[11] Shende D.,Umahiya R.,Raghorte M.,Bhisikar A.,Bhange A.,(2019),"AI Based Voice Assistant Using Python",JETIR February 2019, Volume 6, Issue 2 www.jetir.org (ISSN-2349-

5162).

Legal Assistant Through AI Chatbot In Tamil For Cyber Crimes Against Women

Dr. S.K Lavanya Assistant Professor & *Shriya S, Jayasimman J*

Information Technology Information Technology Information Technology Madras Institute of Technology Madras Institute of Technology Madras Institute of Technology Chennai, India Chennai, India Chennai, India

Tharun CD

Information Technology , Madras Institute of Technology, Chennai, India

In India, awareness of cybercrime has become a necessity, particularly for women who are increasingly targeted by digital offenders. This project explores the development of a Legal Chatbot in the Tamil language, dedicated to empowering women with knowledge about cybercrime laws and legal recourse. The project emphasizes the significance of understanding cybercrime in India, where the number of cases against women has risen alarmingly, as evidenced by an 18.4% surge in cybercrime incidents, with a staggering 28% increase in cases targeting women, according to the National Crime Record Bureau's 2021 report. Existing chatbots, while well-intentioned, often fall short of delivering timely responses, they tend to function merely as intermediaries between victims and legal experts. To address this limitation, our chatbot will deliver precise and pertinent responses, equipping users with the relevant laws, complaint registration procedures, insights from similar cases, and practical guidance. To enhance user interactions, the chatbot will employ advanced NLP models with Named Entity Recognition and Intent identification, a legal knowledge graph, and leverage the power of GPT-3 to generate contextually relevant responses. Furthermore, voice input functionality will be integrated to ensure accessibility to a broader user base. This innovative Legal Chatbot represents a pivotal step in educating and empowering women in the face of rising cybercrimes. By providing immediate, tailored information and support, we aim to make a meaningful contribution to women's safety and legal awareness in the digital age.

AI-Powered YouTube Transcript Summarization with Transformers models

V.Shanmugapriya[1] V.Srividhya[2]
Full-Time Research Scholar[1], Assistant Professor[2]
[1,2] Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, India
Email : [1]sudhamarch22@gmail.com,[2]vidhyavasu@gmail.com

ABSTRACT:

The rapid growth of videos on websites like YouTube in a period of information abundance has created an essential need for effective content summary. This research proposes a novel method for summarizing YouTube transcripts that makes use of state-of-the Transformers models. To make use of artificial intelligence (AI) to generate concise and coherent summaries of wide video transcripts, improving user accessibility and making content easier to read and understand. The background of this work is to shorten the length of the video transcript text. It can be challenging to provide time to watch such videos, which may last longer than expected and viewer's efforts may be useless if they cannot extract useful information from them. Summarizing transcripts of such videos allows us to spot essential patterns in the video and save time quickly. The method of this work consists of four main phases. The first phase is loading the YouTube video. Then apply Preprocessing techniques, which eliminate Punctuation, stop words, and case formatting. Next is to implementation of Abstractive Summarization using Pretrained Transformer models. Bidirectional Auto Regressive Transformers (BART) and Pre-training with Extracted Gap-sentences(PEGASUS) are these two models used for summarization in this particular instance. Next phase is to convert the summarized text into regional language(Tamil).The final of this work is Performance Evaluation using ROUGE Score(Accuracy, precision, recall, and F-measure) to find the best model. The finding is that the pre-trained language models built on the transformer architecture were best suited for summarization tasks. To conduct comparative studies, this is calculated by ROUGE scores for each model's predictions.

Keywords: Abstractive Summarization, Pretrained Transformer-BART Model, Pretrained Transformer - PEGASUS Model, Performance Evaluation.

Code-Mixed "Computationally Romba Challengingaa Irukku"

Kathiravan Pannerselvam[1] and Saranya Rajiakodi[2]
[1,2]Department of Computer Science
Central University of Tamil Nadu-Thiruvarur

Abstract

Code-mixing, the amalgamation of multiple languages in conversation or text, poses unique challenges for natural language processing (NLP). Bilingual speakers seamlessly switch between languages, especially in the context of Tamil-English code-mixing. It is imperative to address several critical challenges to effectively develop computational tools for processing such code-mixed data. These challenges encompass addressing the scarcity of high-quality code-mixed corpora and dealing with privacy concerns during data collection. Furthermore, achieving domain and language generalization in NLP tools and addressing bias and harmful behavior becomes essential. While multilingual models like mBERT and MuRIL offer potential solutions, evaluating their sensitivity to noise and transliterations remains crucial. Standardizing pipelines for code-mixed tasks to enhance performance is of utmost importance. Ongoing research is actively working on prototyping efficient data collection pipelines and conducting in-depth analyses, particularly in the English-Tamil language pair. In conclusion, we require innovative solutions to overcome these challenges and improve NLP tools for code-mixed data, thus benefiting applications such as sentiment analysis, privacy, security, and stance detection.
Keywords- Code-mixed text, Multilingual data, Low resource language, Natural Language Processing, Natural Language Understanding.

1. Introduction

Code-mixing, the dynamic interplay of languages within a single communicative context, has captured the attention of linguists and computational linguists alike. The seamless alternation between languages is a common linguistic practice in multilingual societies worldwide. While code-mixing is an affluent area of linguistic study, this article focuses on its computational aspects, particularly within the Tamil and English context. In Tamil and English code-mixing, the phenomenon takes on unique characteristics due to the coexistence of these languages in various regions, notably in India, Malaysia, Singapore, and Sri Lanka. Tamil, a Dravidian language, and English, a Germanic language, have distinct linguistic features and structures (Kathiravan et al., 2016). The seamless alternation between Tamil and English, influenced by sociolinguistic factors and language policies, poses intriguing challenges and opportunities for natural language processing (NLP) researchers. Code mixing combines linguistic elements, including morphemes, words, modifiers, phrases, clauses, and sentences, combined from two distinct grammatical systems within a single sentence (Report, 2021). Figure 1 depicts the example of Tamil and English code-mixed social media comments.

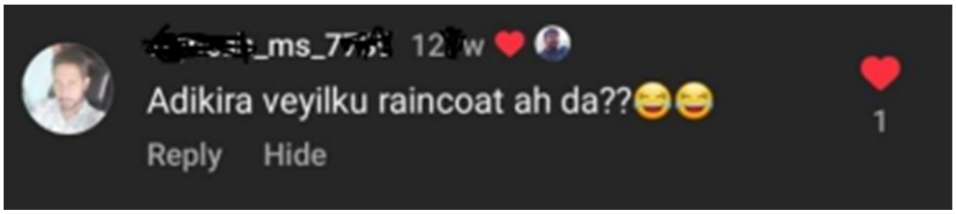


Figure 1: Tamil and English code-mixed text

This research article aims to provide a comprehensive overview of the challenges and prospects in processing code-mixed data involving Tamil and English. Our objectives include: 1. Investigating the scarcity of high-quality code-mixed corpora and privacy concerns in data collection. 2. Exploring the challenges of generalization across domains and languages, particularly in social media text. 3. Analyzing the role of transfer learning and multilingual models in addressing the low-resource nature of code mixed data.
4. Discuss strategies to address bias and harmful behavior in code-mixed data processing. 5. Highlighting the need for standard pipelines and integrated approaches in code-mixed NLP research.
The objectives, as mentioned above, generate the following research questions.

RQ1- What strategies can effectively address the scarcity of high-quality code-mixed corpora and privacy concerns in data collection?

Ans: To address these challenges, researchers can explore alternative data sources like comments on newspaper articles and video-sharing platforms, identify high-yield code-mixed terms within existing corpora, and develop efficient data collection pipelines that respect privacy regulations.

RQ2- How can generalization challenges across domains and languages, especially in social media text, be mitigated?

Ans: Overcoming these challenges involves deeply understanding hashtags' role in encoding affective and semantic content in tweets, leveraging hashtag segmentation tools, and minimizing reliance on quick data annotation methods. Developing models that are not domain-specific or dataset-specific is also crucial.

RQ3- What are the contributions and limitations of transfer learning and multilingual models in addressing the low-resource nature of code-mixed data?

Ans: Transfer learning and multilingual models, like mBERT and MuRIL, offer promise in enriching code-mixed data representations. However, their sensitivity to noise, spelling variations, and transliterations must be critically evaluated to harness their full potential. They are preciousfor low-resource language pairs, such as English-Telugu and English-Kannada, offering solutions for code-mixed text processing.

This research article is structured as follows: Section 2 provides an in-depth discussion of the challenges related to Data Collection and Privacy Concerns. Section 3 explores the topic of Generalization Across Domains and Languages. Section 4 delves into Transfer Learning and Multilingual Models. Section 5 addresses the critical issue of Addressing Bias and Harmful Behavior. Section 6 highlights the Lack of Standard Pipelines in code-mixed NLP. Section 7

presents an overview of Current Work in Progress in the field. Finally, Section 8 offers a Conclusion that discusses the potential impact of code-mixed NLP research. With these objectives and structure in mind, we explore the challenges and solutions associated with processing code-mixed data involving Tamil and English.

Table 1: List of Code-mixed language datasets, sources, and applications

Name	**Languages**	**Source**	**Applications**
Offenseval Dravidian (Chakravarthi et al., 2021)	English-Tamil, English-Malayalam, English - Kannada	YouTube	Offensive language detection
FIRE 2020 Dravidian Code Mixed (Chakravarthi et al., 2021)	Tamil, Malayalam	YouTube	Sentiment Analysis
FIRE 2013-16 Tasks (Banerjee et al., 2020)	English, Hindi, Tamil, Telugu,	Tweets, Facebook,	Transliterated Search, Question Answering
Stance Detection (Srinidhi Skanda et al., 2017)	English – Kannada	Facebook	Stance detection

2. Data Collection and Privacy Concerns

One of the fundamental challenges in code-mixed NLP research is acquiring high-quality data. Traditional sources of text data, such as news articles and Wikipedia, often lack code-mixed content, necessitating innovative approaches to data collection. Speech corpora, text messages, and online social networks have emerged as valuable repositories of code-mixed text. Nevertheless, each of these sources presents its own set of challenges.

Social network companies are under increasing pressure to protect user data, making accessing and collecting data from these sources progressively tricky. The Privacy concerns encompass user consent, data anonymization, and ethical considerations (Banerjee et al., 2020). As researchers, we are responsible for balancing the pursuit of

valuable code-mixed data with the need to protect user privacy. Striking this balance requires careful consideration and ethical data collection practices. Table 1 illustrates the Code-mixed

datasets and their applications with the data sources.

Researchers are exploring innovative data collection approaches to overcome the scarcity of code-mixed corpora and navigate the privacy challenges. These approaches include:

1. Identifying high-yield code-mixed terms within existing corpora to augment code-mixed data.

2. Exploring unconventional sources such as comments on newspaper articles and video-sharing platforms, where code-mixing is prevalent.

3. Leveraging open-source initiatives and community contributions to build code-mixed datasets collaboratively.

These approaches address data scarcity and promote community and collaboration among researchers in the code mixed NLP field (Chakravarthi et al., 2022; Kumaresan et al., 2021). Researchers must remain vigilant about privacy concerns and adopt responsible data collection practices to collect code-mixed data involving Tamil and English. Balancing the need for data with ethical considerations is essential for advancing code-mixed NLP.

3. Generalization Across Domains and Languages

Developing NLP tools to process code-mixed data effectively requires models that generalize across different domains and languages. However, the dynamic and diverse nature of textual data on social media platforms presents unique challenges to achieving this level of generalization. Social media platforms are dynamic environments where users discuss various topics. Code-mixed text on these platforms spans various domains, from politics and entertainment to sports and personal narratives. Models trained on a specific domain may struggle to generalize to new and diverse domains, impacting their usability across different types of code-mixed data. To address this challenge, researchers must develop robust and adaptable models to domain shifts. This requires the incorporation of diverse training data and strategies for domain adaptation.

Code-mixed text exhibits multifaceted linguistic characteristics. It can include elements of both languages, such as vocabulary, grammar, and syntax, seamlessly interwoven. Moreover, colloquial language, slang, and cultural references further complicate the analysis of code-mixed data. Understanding and accommodating these linguistic nuances is crucial for accurate processing. Tools that effectively separate the linguistic features of code-mixed text and recognize the switching points between languages are pivotal for accurate NLP in this context.

On social media platforms, hashtags serve as a means of categorization and expression. They play a vital role in encoding affective and semantic content in code-mixed tweets and posts. Understanding the significance of hashtags and leveraging hashtag segmentation tools can significantly impact the development of tools for code mixed data and social media text analysis. By recognizing the emotional and semantic cues conveyed through hashtags, NLP models can enhance their understanding of code-mixed text and improve the accuracy of sentiment analysis and content classification.

Achieving robust generalization across diverse domains and effectively deciphering the multifaceted nature of code-mixed text are pivotal challenges in developing NLP tools for code mixed data involving Tamil and English. Moreover, recognizing the importance of hashtags as linguistic markers can unlock new avenues for enhancing NLP accuracy in this context.

4. Transfer Learning and Multilingual Models

The scarcity of code-mixed data, particularly for low-resource language pairs like Tamil and English, necessitates innovative approaches to enriching representations. Multilingual models have emerged as a promising solution for code-mixed data processing, but their sensitivity to noisy text, spelling variations, and transliterations must be critically evaluated(Antoun et al., 2020; Feng et al., 2022; Kalaivani et al., 2021). Code-mixed data, especially for less commonly studied language pairs like Tamil and English, is often characterized by insufficient training data. This low-resource nature poses a significant challenge for traditional NLP approaches that rely on large, well annotated corpora. Multilingual models, such as mBERT (Multilingual BERT), have demonstrated great promise in cross-lingual model transfer (G. K. Kumar et al., 2022). These models can be fine-tuned and evaluated for various languages, including code-mixed language pairs. However, their effectiveness centers on carefully handling linguistic variation, including noisy text, spelling variations, and transliterations. Researchers must critically assess the suitability of multilingual models for code-mixed data, identifying potential pitfalls and limitations in their application. Recent developments have introduced models like MuRIL (Multilingual
Representations for Indian Languages). These models are trained on monolingual corpora of Indian languages and their transliterated counterparts. MuRIL offers a promising solution for code-mixed text processing involving Indian languages, including Tamil and English (S. Kumar et al., 2022). Transfer learning and multilingual models offer a lifeline for code-mixed data processing in low-resource scenarios. However, researchers must navigate linguistic variation challenges and critically evaluate these models' suitability for specific code-mixed language pairs, such as Tamil and English.

5. Addressing Bias and Harmful Behavior

The deployment of large language models for code-mixed text processing, often sourced from online social networks, introduces concerns regarding harmful behavior and bias. Mitigating these effects is crucial for responsible NLP research. While powerful, large language models are not immune to exhibiting toxic and biased behavior, especially when trained on web data that includes user-generated content. Code-mixed research often relies on data from online social networks, which may contain hate speech, offensive language, or biased content. Mitigating the harmful effects of deploying such models is paramount, necessitating rigorous evaluation and safeguards (Chakravarthi et al., 2022; Ghanghor et al., 2021; Kumaresan et al., 2021; Ravikiran et al., 2022). Code-mixed data processing adds a layer of complexity to the assessment of model behavior. Researchers must characterize the harmfulness of models in the code-mixed setting, taking into account linguistic nuances, cultural sensitivities, and sociolinguistic factors specific to the

language pair. This characterization involves the development of evaluation metrics and guidelines tailored to code-mixed contexts (Mahmud et al., 2023; Sanh et al., 2019).

Ethical considerations and responsible research practices are foundational in code-mixed NLP. Researchers must adhere to ethical data collection, model training, and evaluation guidelines. Additionally, they should actively engage with communities affected by code-mixed language practices to ensure culturally sensitive and unbiased research. Transparency, fairness, and ethical conduct should be at the forefront of code-mixed NLP research (Nascimento et al., 2022). Addressing bias and harmful behavior in code-mixed NLP is a technical challenge and an ethical imperative. Researchers must navigate these concerns carefully and diligently to ensure the responsible development and deployment of NLP tools.

6. Lack of Standard Pipelines

Efficient and effective NLP tools for code-mixed data require standardized pipelines integrating essential steps, such as language identification (LID) and normalization. The absence of such pipelines can hinder performance on complex NLP tasks. Code-mixed NLP research often involves a fragmented landscape, where various tasks and associated datasets are treated in isolation. This fragmentation can lead to suboptimal performance on more complex NLP tasks, as essential components like language identification and normalization are often overlooked or inconsistently applied. Streamlining and unifying these processes within a standard pipeline can improve overall NLP performance for code-mixed data.

6.1 The Role of Language Identification (LID)

Language identification, the process of determining the language(s) present in a text, is a fundamental step in code-mixed data processing. Accurate LID is crucial for applying language-specific NLP tools and models effectively. Integrating LID into the code-mixed NLP pipeline ensures that subsequent processing steps are tailored to the identified languages (Hidayatullah et al., 2022). Developing robust and language-agnostic LID models is a key component of standardized code-mixed NLP pipelines.

6.2 Normalization and Preprocessing

Code-mixed text often requires normalization and preprocessing to address linguistic variations, spelling inconsistencies, and other linguistic complexities. Normalization ensures that text is in a standardized format for downstream NLP tasks. Establishing standardized normalization techniques within code-mixed NLP pipelines can enhance the reliability of these tools (Thara & Poornachandran, 2018). Normalization methods may vary depending on the language pair and the nature of the code-mixed text, requiring flexibility within the pipeline.

The pipelines in code-mixed NLP research is a hurdle that must be overcome. Creating unified pipelines encompassing essential steps such as language identification and normalization will facilitate collaboration, improve performance, and drive progress in the code-mixed NLP community.

7. Current Work in Progress

Code-mixed NLP research is an evolving field, with ongoing efforts to address the challenges discussed in this article. This section highlights some of the latest developments and research endeavors.

7.1 Efficient Data Collection and Annotation

Ongoing research efforts focus on developing efficient pipelines for collecting code-mixed corpora. These pipelines incorporate advanced techniques such as sentence-level classification and query term mining for social media APIs. This approach streamlines the process of acquiring code-mixed data for research and development. Efforts are underway to leverage machine learning and natural language processing techniques to automate aspects of data collection and annotation, reducing the manual effort required.

7.2 Quantitative and Qualitative Analysis of Code-Mixed Utterances

Researchers conduct in-depth quantitative and qualitative analyses of code-mixed utterances, particularly in language pairs like English and Tamil. These analyses encompass various linguistic dimensions, including parts of-speech tagging, syntactic analysis, and semantic interpretation. Researchers aim to develop more accurate and context-aware NLP tools by gaining a deeper understanding of code-mixed text. Advancements in linguistic analysis tools and methodologies drive progress in understanding code-mixed language practices.

7.3 Community Engagement and Collaboration

Collaboration and community engagement are at the heart of code-mixed NLP research. Researchers are actively collaborating with linguists, language communities, and experts in sociolinguistics to ensure culturally sensitive and context-aware research. This collaborative approach is instrumental in addressing code-mixed data's linguistic and sociocultural complexities. Community-driven initiatives, data sharing, and open-source development are contributing to the growth of the code-mixed NLP research ecosystem. As code-mixed NLP research continues to evolve, these ongoing efforts reflect the commitment of researchers to tackle the challenges posed by code-mixing involving Tamil and English. These endeavors aim to enhance the usability and reliability of NLP tools for understanding and analyzing code-mixed text, ultimately benefiting various applications, including sentiment analysis, privacy and security, and stance detection.

8. Discussion and Conclusion

Code-mixing, the dynamic interplay of languages within a single communicative context, presents unique challenges and opportunities for natural language processing (NLP) researchers. In the context of Tamil and English, bilingual speakers seamlessly switch between these languages, creating a rich linguistic landscape. To harness the potential of code-mixed data involving Tamil and English, researchers must navigate challenges in data collection, model development, and ethical considerations. Addressing the scarcity of high-quality code mixed corpora requires innovative solutions. Researchers are exploring unconventional data sources, identifying high-

yield code-mixed terms within existing corpora, and fostering community-driven initiatives. These efforts aim to balance data needs with ethical considerations surrounding user privacy.

Generalizing NLP tools across diverse domains and languages, particularly in social media text, remains a formidable challenge. Researchers must develop models that adapt to domain shifts and recognize the multifaceted nature of code-mixed text. Additionally, understanding the role of hashtags as linguistic markers holds promise for enhancing NLP accuracy. Transfer learning and multilingual models offer a lifeline for code-mixed data processing in low-resource scenarios. However, their sensitivity to linguistic variations and transliterations necessitates careful evaluation. Tailored models, like MuRIL, demonstrate the value of language-specific approaches to code-mixed text.

Addressing bias and harmful behavior in code-mixed NLP is an ethical imperative. Researchers must characterize model harmfulness in code-mixed settings and adhere to ethical data collection and research practices. Engaging with affected communities is crucial for culturally sensitive and unbiased research. Standardized pipelines that encompass language identification, normalization, and other essential steps are essential for efficient and effective code-mixed NLP. Streamlining these processes promotes collaboration and improves performance in code-mixed NLP research. The code-mixed NLP research community is actively
working to address these challenges. Efficient data collection, linguistic analysis, community collaboration, and open-source development are driving progress in the field.

In conclusion, the journey of processing code-mixed data involving Tamil and English is marked by challenges and prospects. As researchers, we are dedicated to overcoming these challenges through innovation, ethical research practices, and collaborative efforts. By doing so, we aim to enhance the usability and reliability of NLP tools for understanding and analyzing code-mixed text. These advancements will ultimately benefit various applications, including sentiment analysis, privacy and security, and stance detection, and contribute to a deeper understanding of multilingual communication in the digital age.

References

Antoun, W., Baly, F., & Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*. http://arxiv.org/abs/2003.00104

Banerjee, S., Choudhury, M., Chakma, K., Naskar, S. K., Das, A., Bandyopadhyay, S., & Rosso, P. (2020). MSIR@FIRE: A Comprehensive Report from 2013 to 2016. *SN Computer Science*, *1*(1). https://doi.org/10.1007/s42979-019-0058-0

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2022). DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*, *56*(3), 765–806. https://doi.org/10.1007/S10579-022-09583-7

Chakravarthi, B. R., Priyadharshini, R., Thavareesan, S., Chinnappa, D., Thenmozhi, D., Sherly, E., McCrae, J. P., Hande, A., Ponnusamy, R., Banerjee, S., & Vasantharajan, C. (2021). Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text. *CEUR Workshop Proceedings*, *3159*, 872–886.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 878– 891. https://doi.org/10.18653/v1/2022.acl-long.62

Ghanghor, N. K., Krishnamurthy, P., Thavareesan, S., Priyadarshini, R., & Chakravarthi, B. R. (2021). IIITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada. *Proceedings of the 1st Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech 2021 at 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*, 222–229.

Hidayatullah, A. F., Qazi, A., Lai, D. T. C., & Apong, R. A. (2022). A Systematic Review on Language Identification of Code-Mixed Text: Techniques, Data Availability, Challenges, and Framework Development. *IEEE Access*, *10*(November), 122812–122831. https://doi.org/10.1109/ACCESS.2022.3223703

Kalaivani, A., Thenmozhi, D., & Aravindan, C. (2021). TOLD: Tamil Offensive Language Detection in Code Mixed Social Media Comments using MBERT with Features based Selection. *CEUR Workshop Proceedings*, *3159*, 667–679.

Kathiravan, P., Makila, S., Prasanna, H., & Vimala, P. (2016). Over view—The machine translation in NLP. *Int. J. Sci. Adv. Technol*, *2*(7), 19–25.

Kumar, G. K., Gehlot, A. S., Mullappilly, S. S., & Nandakumar, K. (2022). MuCoT: Multilingual Contrastive Training for Question-Answering in Low-resource Languages. *DravidianLangTech 2022 - 2nd Workshop on Speech and Language Technologies for Dravidian Languages, Proceedings of the Workshop*, *1*, 15–24. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.3

Kumar, S., Rajesh, D. D., Pranesh, S., Kollipara, V. N. H., Agrawal, G. K., Anbarasi, M., & J, V. (2022). Classification of Indian media titles using deep learning techniques. *International Journal of Cognitive Computing in Engineering*, *3*, 114–123. https://doi.org/10.1016/J.IJCCE.2022.04.001

Kumaresan, P. K., Premjith, Sakuntharaj, R., Thavareesan, S., Navaneethakrishnan, S., Madasamy, A. K., Chakravarthi, B. R., & McCrae, J. P. (2021). Findings of Shared Task on Offensive Language Identification in Tamil and Malayalam. *ACM International Conference Proceeding Series*, 16–18. https://doi.org/10.1145/3503162.3503179

Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing and Management*, *60*(5), 103454. https://doi.org/10.1016/j.ipm.2023.103454

Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, *201*, 117032. https://doi.org/10.1016/J.ESWA.2022.117032

Ravikiran, M., Chakravarthi, B. R., Madasamy, A. K., Sivanesan, S., Rajalakshmi, R., Thavareesan, S., Ponnusamy, R., & Mahadevan, S. (2022). Findings of the Shared Task on Offensive Span Identification from Code-Mixed Tamil-English Comments.

DravidianLangTech 2022 - 2nd Workshop on Speech and Language Technologies for Dravidian Languages, Proceedings of the Workshop, 261–270. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.40

Report, C. V. (2021). *Code Mixing " computationally bahut challenging hai ." August*, 1–37.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108

Srinidhi Skanda, V., Anand Kumar, M., & Soman, K. P. (2017). Detecting stance in kannada social media code mixed text using sentence embedding. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-January*, 964–969. https://doi.org/10.1109/ICACCI.2017.8125966

Thara, S., & Poornachandran, P. (2018). Code-Mixing: A Brief Survey. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, 2382–2388. https://doi.org/10.1109/ICACCI.2018.8554413

Deep Learning for Sarcasm Identification in Tamil-English Code-mixed Data

Ramya Priya S[1], Shanmitha Thirumoorthy[2], DurairajThenmozhi
[1] Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India.
[2] Vellore Institute of Technology, Chennai, Tamil Nadu, India
E-mail(s): ramyapriya19088@cse.ssn.edu.in; shanmitha.t2023@vitstudent.ac.in; theni_d@ssn.edu.in

Abstract

This paper presents a novel approach to sarcasm identification in Dravidian languages, specifically Tamil, focusing on code-mixed text commonly found in social media platforms. Sarcasm, a complex linguistic expression, often challenges traditional language rules by conveying the opposite of its literal interpretation. Recognizing sarcasm is crucial for accurate sentiment analysis and understanding the underlying intent and context in multilingual environments. We propose a sequential neural network model that processes code-mixed Tamil-English text collected from social media. Our pre-processing techniques include language detection and transliteration to ensure consistency. The model employs embedding layers, dense layers with ReLU activation, and dropout layers to capture intricate text patterns and prevent overfitting. The final layer, utilizing sigmoid activation, facilitates binary classification for sarcasm detection. Evaluation on the FIRE 2023 test dataset demonstrates a commendable overall accuracy of 78%, with a precision of 84% for non-sarcastic statements and a recall of 90%. The macro average F1-score is 0.72, and the weighted average F1-score is 0.78, emphasizing the model's balanced and robust performance. These findings highlight the model's potential utility in sentiment analysis, customer feedback analysis, and content moderation in Tamil language platforms, contributing to a deeper understanding of sarcasm in Tamil.

Keywords: Sarcasm Identification, Code-Mixed text, Neural Network, Tamil, Natural Language Processing

1 Introduction

Sarcasm, a sophisticated form of linguistic expression where the intended meaning of a statement is contrary to its literal interpretation, poses a formidable challenge in natural language understanding. Detecting sarcasm in written text, including social media posts, comments, reviews, and news articles, has emerged as a critical task in the realm of natural language processing. This task hinges on the automatic recognition of sarcastic statements within a given corpus of text, which, in turn, is pivotal for comprehending the genuine sentiment, intent, and context concealed within a statement. Sarcasm, by its nature, often subverts conventional language rules, necessitating advanced computational techniques to uncover its subtleties. The significance of sarcasm identification spans across diverse domains and applications. In the realm of sentiment analysis, for instance, accurately discerning sarcastic comments from their non-sarcastic counterparts is indispensable for precisely categorizing sentiment.

One emerging challenge in the field of sarcasm identification is the growing demand for effective

methods on social media texts, especially in the context of Dravidian languages like Tamil. A distinctive characteristic of social media communication in these languages is code-mixing, a prevalent phenomenon where multiple languages are combined within a single discourse. Code-mixed texts are often written in non-native scripts, further complicating the task. Systems trained on monolingual data tend to falter when faced with the intricacies of code-switching across different linguistic levels in such texts. Recognizing the urgency of addressing this challenge, our research focuses on sarcasm identification in Tamil, leveraging a sequential neural network model, and employing a code-mixed dataset comprising comments and posts collected from social media platforms. This dataset encompasses both Tamil and English, adding a layer of complexity and nuance to the sarcasm detection task. Our approach involves extensive pre-processing, including language detection and transliteration to ensure uniformity in the data. Subsequently, the data is tokenized into sequences of word indices, ready for neural network processing.

In the subsequent sections, we delve into the intricacies of our methodology, detailing the architecture of our sequential neural network, the pre-processing steps, and the evaluation metrics. We present the results of our model on the Tamil-English Dravidian-CodeMix dataset, showcasing its exemplary performance. Our findings underscore the potential utility of our model in applications like sentiment analysis, customer feedback analysis, and content moderation in Tamil language platforms, ultimately contributing to a more comprehensive understanding of sarcasm in the Tamil language, a promising facet of natural language processing that holds immense prominence and relevance in today's digital age.

2 Related Work

This section provides an overview of the existing approaches to identify sarcasm. The majority of research in sarcasm identification has primarily focused on the English language, given its prevalence in social media communication. Recent studies have made significant strides in identifying sarcasm within English-scripted domains like Twitter, product reviews, website comments, and others, as evidenced by numerous research efforts[12-14]. However, in the context of low-resource languages such as Hindi, Telugu, Tamil, Chinese, Arabic, and others, there has been relatively limited exploration. Subsequent subsections will delve into the specifics of sarcasm detection in both English and low-resource languages[15-16].

Sarcasm Detection in the English Language

Early research in sarcasm detection often employed rule-based methods, relying on linguistic patterns and lexico-syntactic cues. For instance, Riloff et al. (2013)[1] introduced a rule-based approach that identified sarcasm by contrasting positive sentiment with negative situations. Supervised machine learning techniques, including Support Vector Machines (SVMs) and Random Forests, gained prominence for sarcasm detection. Researchers like Davidov et al. (2010)[2] and Reyes et al. (2013)[3] utilized labeled datasets to train models that could identify sarcastic utterances based on features such as n-grams and sentiment scores. (Das, D., & Clark, A. J. ,2018)[4] presented an approach based on supervised machine learning considering posts having text and images as content and also user's interaction on those Facebook posts to detect sarcasm. They proposed that if multimedia contents (like images) also shared along with textual posts then that can prove useful in detecting sarcasm. With the rise of deep learning, neural network-based

models have been increasingly employed for sarcasm detection. Some studies have utilized recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks to capture sequential dependencies in text (Gonçalves et al., 2018)[5]. Current approaches to automatic sarcasm detection rely primarily on lexical and linguistic cues. Ashwin et al aims to address the difficult task of sarcasm detection on Twitter by leveraging behavioral traits intrinsic to users expressing sarcasm.

Avinash et al [7] introduced a multi-head attention-based bidirectional long-short memory (MHA-BiLSTM) network to detect sarcastic comments in a given corpus. They extract the most significant features and build a feature-rich SVM that outperforms models built using lexical, semantic and pragmatic features. (Gupta, S., Singh, R., & Singla, V., 2020) [6] proposed sarcasm detection system based on emoticons and text. Further for detecting both sarcasm and emoticons, two polarities were identified that is positive and negative, thus achieving an accuracy of 100%.They used artificial neural network (ANN) as a classifier to classify the polarities. Also to increase the emoticon polarity detection emoji sentiment ranking lexicon detection
system was used.

Sarcasm detection in the English language has witnessed a progression from rule-based approaches to sophisticated deep learning models. While much of the research has been focused on English, there is an increasing interest in multilingual sarcasm detection, as sarcasm is not limited to a single language. This opens avenues for research in languages other than English.

Sarcasm Detection on Low Resourced Languages

Sarcasm identification in low-resource languages is a relatively uncharted territory within the field of natural language processing (NLP). Such languages, characterized by limited linguistic resources, pose distinct challenges for building robust sarcasm detection models due to the scarcity of large annotated datasets essential for training. While the literature primarily focuses on well-resourced languages, recent studies in related languages offer valuable insights and methodologies applicable to Dravidian languages like Tamil.

Akshi et al. [8] made significant strides in the domain by presenting a Hybrid Deep Learning Model for Sarcasm Detection in Indian Indigenous Languages, with a specific focus on Hindi. Utilizing Word-Emoji Embeddings, their model showcased the pivotal role of emojis in sarcasm detection. Validation on a Hindi tweets dataset, Sarc-H, demonstrated impressive results, achieving an accuracy of 97.35% with an F-score of 0.9708. This research underscores the importance of considering indigenous languages and contextual cues in sarcasm detection.

Deepak et al. [9] proposed a novel approach to sarcasm detection in code-switched tweets, particularly the fusion of English and Indian native language, Hindi. Their hybrid model integrated bidirectional long short-term memory with a softmax attention layer and convolutional neural networks, offering real-time sarcasm detection capabilities. This research extends the applicability of deep learning techniques to multilingual code-switched scenarios, a relevant consideration for Dravidian languages.

Bharti et al. [10] enriched the literature by curating and annotating a corpus of Telugu conversation

sentences, designed explicitly for sarcasm detection. They introduced algorithms based on hyperbolic features, including Interjection, Intensifier, Question mark, and Exclamation symbol, for effective sarcasm analysis in Telugu conversation sentences. This work contributes to understanding sarcasm detection nuances in Dravidian languages by focusing on Telugu.

Furthermore, a deep learning-based approach [11] addressed the challenge of sarcasm detection in Hindi-English code-mixed tweets. This study leveraged bilingual word embeddings derived from FastText and Word2Vec approaches, emphasizing the importance of bilingual resources for multilingual sarcasm detection. By targeting code-mixed text, this research bridged a gap in comprehending sarcasm in multilingual contexts.

It is noteworthy that, as of our knowledge cutoff date, no specific work on sarcasm detection in Tamil, a prominent Dravidian language, has been reported. Given the linguistic diversity and unique characteristics of Dravidian languages, there is a promising avenue for future research to explore sarcasm detection in languages like Tamil, potentially building upon the insights and methodologies offered by related studies in other low-resource languages. This gap underscores the need for further investigation into sarcasm detection within the Dravidian language family.

3 Proposed Methodology

This proposed methodology combines word embeddings, deep learning, and sequential neural network architecture to address sarcasm identification in Tamil text data. It begins with data pre-processing, followed by the definition and compilation of the neural network model. The model is then trained and validated to achieve the desired sarcasm detection accuracy.

Dataset

The Dravidian – CodeMix – FIRE 2023 Tamil – English dataset was used for training and evaluation of the sarcasm identification model.It is a code-mixed dataset of comments/posts in Tamil-English collected from Youtube Video Comments. A comment/post may contain more than one sentence, but the average sentence length of the corpora is 1. Each comment/post is annotated with sentiment polarity at the comment/post level. This dataset also has class imbalance problems depicting real-world scenarios. The dataset contains all three types of code-mixed sentences Inter-Sentential switch, Intra-Sentential switch, and Tag switching. Most comments were written in native script and Roman script with Tamil grammar with English lexicon or English grammar with Tamil lexicon. Each comment is labeled as either sarcastic or non-sarcastic.

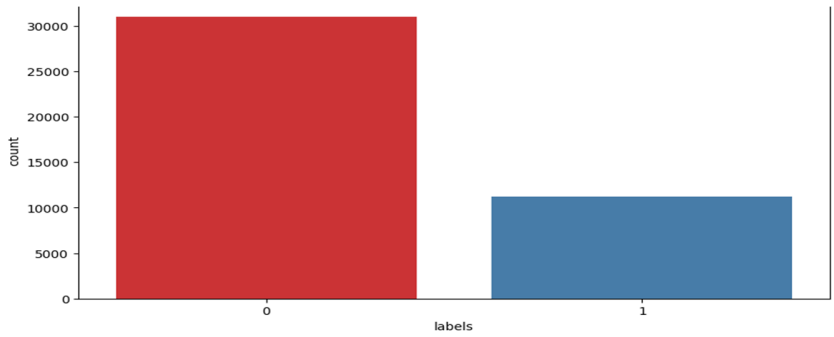


Fig 1. Count plot of Comments

The fig1 shows the count plot of comments labeled as sarcastic and non- sarcastic. For better classification purposes the labels were changed to numbers with 1 representing sarcastic comments and 0 representing non-sarcastic comments.

Data Cleaning and Pre-processing

In the context of sarcasm detection in Tamil text, effective data cleaning and pre-processing are crucial steps to prepare the dataset for model training. This section outlines the data preparation steps undertaken to ensure the quality and relevance of the text data.

Language Detection and Transliteration:

The first step involves identifying the language of each text entry. This is done using the 'langdetect' library, which detects the language of a given text. The goal is to identify Tamil text within the dataset.Detected Tamil text is transliterated from Tamil script to the ITRANS script using the 'indic_transliteration' library. This standardizes the text and ensures uniformity in representation.

Label Mapping:

The labels in the dataset are mapped to numerical values to facilitate model training. In this case, 'Sarcastic' is mapped to 1, and 'Non-sarcastic' is mapped to 0.

Text Cleaning:

A function called 'clean_text' is defined to perform text cleaning. The cleaning steps include:

- Converting text to lowercase to ensure consistency.
- Removing text within square brackets, which often contains nonessential information.
- Removing punctuation marks to focus on text content.
- Eliminating words containing digits, as they may not be relevant for sarcasm detection.
- Removing common English stopwords to reduce noise.

WordCloud Visualization:

A WordCloud is generated to visualize the most frequent words in the cleaned text data. This visualization can provide insights into the prominent words and themes within the dataset. The fig2 and fig3 depict the WordCloud for sarcastic and non-sarcastic comments.

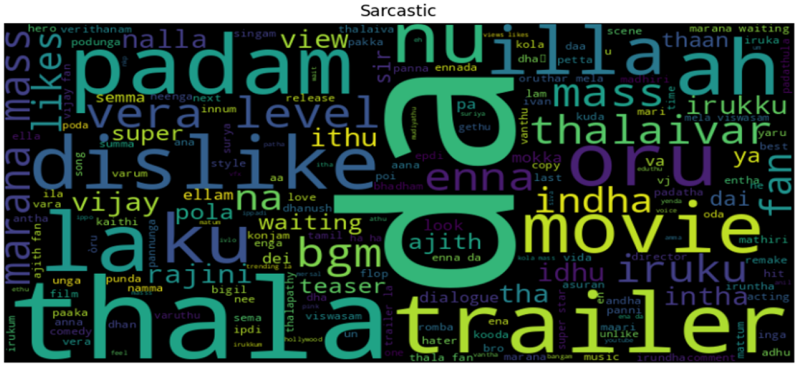


Fig2. Word Cloud for Sarcastic Comments

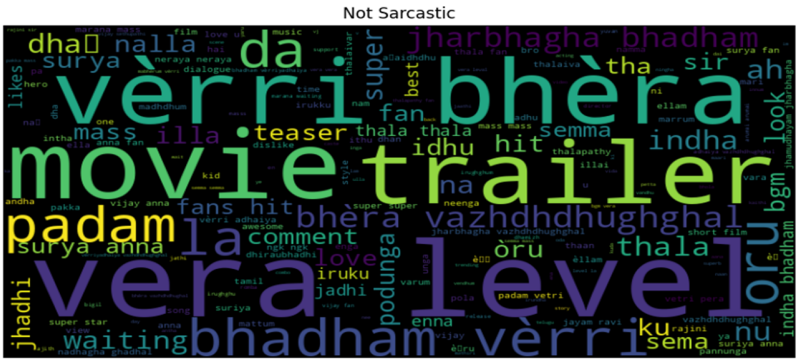


Fig3. Word Cloud for Non-Sarcastic Comments

Tokenization and Padding:

Tokenization is performed using the Keras Tokenizer with specified parameters:

- vocab_size: 10,000 - Limiting the vocabulary size to control the number of unique words.
- oov_tok: '<OOV>' - An out-of-vocabulary token to handle unknown words.

The tokenizer is fitted on the training text data to create a word-to-index mapping.

Training, validation, and test text data are tokenized and padded to a fixed length of 10 words using the 'pad_sequences' function. This ensures that input sequences have consistent lengths.These data cleaning and pre-processing steps are essential for preparing the text data for sarcasm detection model training. The steps not only standardize and clean the text but also enable the conversion of text data into numerical input that can be fed into a neural network for further analysis and sarcasm detection.

Sarcasm Identification Model

The model used is a sequential neural network, constructed using TensorFlow and Keras, designed to effectively identify sarcasm in Tamil text data. The model employs a deep neural network architecture with embeddings, dense layers, and dropout layers to effectively identify sarcasm in Tamil text data. The use of global max-pooling and ReLU activation functions enhances its feature extraction capabilities, while dropout layers help prevent overfitting. The model is well-suited for binary classification tasks and offers a comprehensive solution for sarcasm identification in Tamil language text. The architecture of the neural model is depicted in Fig3.

Below, we outline the key components of our model:

1. Embedding Layer:

The model begins with an Embedding layer, which plays a pivotal role in capturing the semantic meaning of words in the input text.
Parameters:

- vocab_size: 10,000 - This parameter limits the vocabulary size to control the number of unique words considered.
- embedding_dim: 200 - The embedding dimension determines the size of word vectors and helps capture contextual information.
- input_length: 10 - We set a maximum sequence length of 10 words to standardize input data.

2. GlobalMaxPooling1D Layer:

Following the Embedding layer, a GlobalMaxPooling1D layer is applied. This layer extracts the most significant features from the sequence of word embeddings.
Global max-pooling helps reduce dimensionality while retaining critical information.

3. Dense Layers (Three Layers):

Our architecture incorporates three fully connected Dense layers, each introducing non-linearity and abstracting complex patterns.
Configuration of Dense layers:
First Dense Layer: 40 neurons with ReLU activation.
Second Dense Layer: 20 neurons with ReLU activation.
Third Dense Layer: 10 neurons with ReLU activation.

4. Dropout Layers (Three Layers):

To mitigate overfitting, Dropout layers are strategically placed after each Dense layer.Dropout randomly deactivates a fraction of neurons during training, enhancing the model's generalization.

5. Final Dense Layer:

The model concludes with a Dense layer consisting of a single neuron.This neuron uses a sigmoid activation function, making it suitable for binary classification. The output represents the model's prediction for sarcasm, with 0 indicating non-sarcastic and 1 indicating sarcastic.

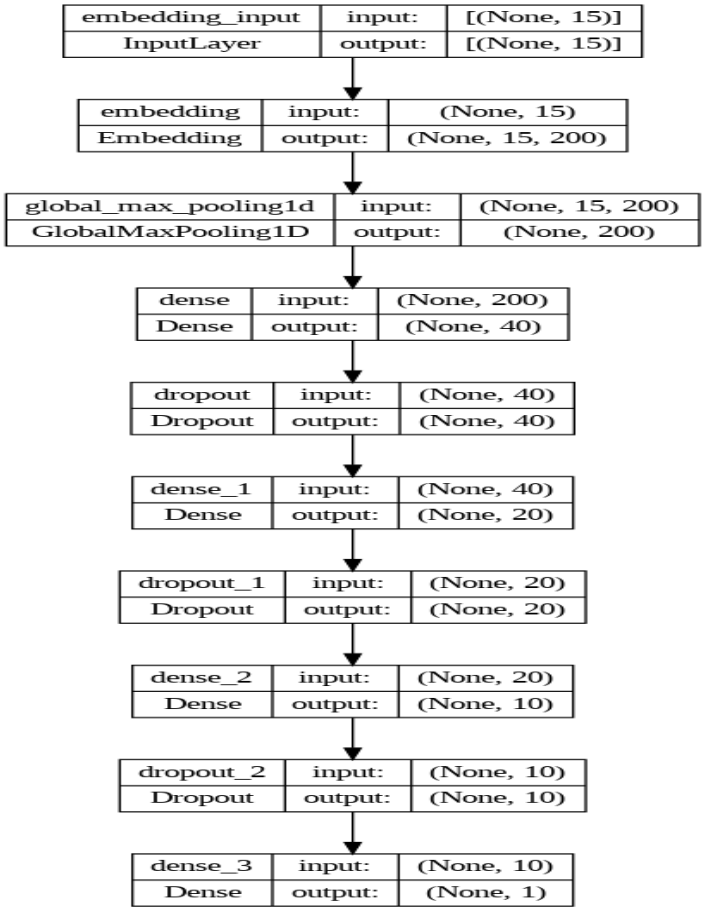


Fig3. Neural Network Architecture

Model Compilation:

During compilation, the following parameters are specified:
Loss Function: Binary Cross-Entropy - A well-suited choice for binary classification tasks.
Optimizer: Adam - An adaptive optimization algorithm that adjusts learning rates during training.
Evaluation Metric: Accuracy - Used to assess the model's performance.

Training:
The model is trained for a predefined number of epochs (in this case, 5 epochs).
Training data, including tokenized and padded sequences, are used to update the model's internal parameters.

4 Experimental Results

The results of our sarcasm identification model evaluated on the Dravidian – CodeMix – FIRE 2023 Tamil – English dataset are presented in the following classification report. There are three statistical parameters namely, Precision, Recall and F –score used to evaluate the proposed approaches. Precision shows how much relevant information is identified correctly and Recall shows how much extracted information is relevant. F – score is the harmonic mean of Precision and Recall.
Equations below show the formula to calculate Precision, Recall and F – score.

$$\text{Precision} = Tp / Tp + Fp$$

$$\text{Recall} = Tp / Tp + Fn$$

$$F - \text{Score} = (2 * \text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$$

where, Tp = True Positive, Fp = False Positive, Fn = False Negative.

The model was evaluated using a comprehensive set of metrics, including precision, recall, and the F1-score, to assess its performance in distinguishing between sarcastic and non-sarcastic statements. The results are shown in Table 1.

	Precision	Recall	F1-Score	Support
Not Sarcastic	0.86	0.84	0.85	3097
Sarcastic	0.59	0.62	0.60	1128
Accuracy			**0.78**	**4225**
Macro Avg	**0.72**	**0.73**	**0.73**	**4225**
Weighted Avg	**0.79**	**0.78**	**0.78**	**4225**

Table 1. Classification Report

Our model achieved an overall accuracy of 78%, demonstrating its ability to accurately classify sarcasm in Tamil text. Notably, the model exhibited strong precision in identifying non-sarcastic statements at 86%, indicating a low rate of false positives. The recall for non-sarcastic statements was 84%, indicating that the model effectively captured the majority of non-sarcastic instances. In case of sarcastic statements, the model achieved a precision of 59%, implying that it correctly classified 59% of statements as sarcastic out of the total classified as sarcastic. The recall for sarcastic statements was 62%, indicating that the model identified 62% of the sarcastic statements present in the dataset.

The macro average F1-score, which balances precision and recall across both classes, is 0.73, reflecting a balanced performance in sarcasm detection. The weighted average F1-score is 0.78, highlighting the model's robust overall performance.

The confusion matrix for our sarcasm identification model on the Dravidian – CodeMix – FIRE 2023 Tamil – English test dataset is presented in fig5. This matrix provides a detailed breakdown

of the model's performance in classifying statements as sarcastic or non-sarcastic. The model correctly identified 2065 non-sarcastic statements as not sarcastic, demonstrating its ability to accurately classify non-sarcastic text.The model correctly identified 700 sarcastic statements as sarcastic, showcasing its effectiveness in identifying sarcasm in Tamil text.

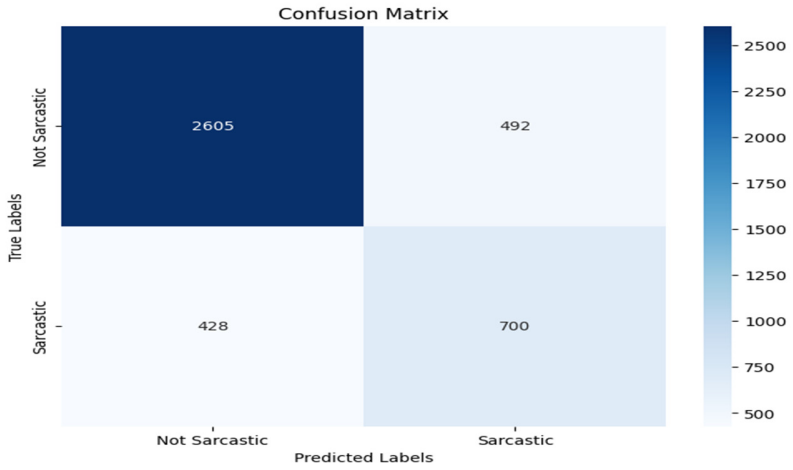


Fig5. Confusion Matrix

These results underscore the effectiveness of our sequential neural network model in identifying sarcasm in Tamil text. The model's performance is promising and holds significant potential for various applications, including sentiment analysis, customer feedback analysis, and content moderation in Tamil language platforms.

5 Conclusions

Sarcasm identification in Dravidian language Tamil, particularly in code-mixed data, is a challenging and crucial task in the realm of natural language processing. In this study, we addressed this challenge by developing a sequential neural network model designed to discern between sarcastic and non-sarcastic statements in Tamil text, even in the presence of code-mixing. Our model achieved an overall accuracy of 78%, showcasing its ability to accurately classify sarcasm in Tamil text. The macro average F1-score of 0.73 showcases a balanced performance across both classes. The weighted average F1-score of 0.78 underscores the model's robust overall performance.

Our research holds significant implications for various applications, including sentiment analysis, customer feedback analysis, and content moderation on Tamil language platforms. Accurate sarcasm identification is pivotal for understanding sentiment and context, contributing to a more comprehensive understanding of user-generated content. In conclusion, our study contributes to the emerging field of sarcasm detection in Dravidian languages, providing a valuable foundation for future endeavors. Our sequential neural network model, designed to handle code-mixed data, showcases potential for a wide range of applications, ultimately advancing the understanding of

sarcasm in the Tamil language.

References

[1] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 704-714.

[2] Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010), 107-116.

[3] Reyes, A., Rosso, P., & Veale, T. (2013). A Multilingual Lexicon of Sarcasm Built through Crowdsourcing. Language Resources and Evaluation, 47(3), 1005-1028.

[4] Das, D., & Clark, A. J. (2018, October). Sarcasm detection on Facebook: A supervised learning approach. In Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct (pp. 1-5)

[5] Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2018). Sarcasm Detection in Microblogs: A Multilingual Corpus-Based Approach. ACM Transactions on the Web (TWEB), 12(4), 1-28.

[6] Gupta, S. , Singh, R., & Singla, V. (2020) Emoticon and Text Sarcasm Detection in Sentiment Analysis. In First International Conference on Sustainable Technologies for Computational Intelligence (pp. 1-10), Springer, Singapore.

[7] Gella, S., Lewis, M. and Rohrbach, M., 2018. A dataset for telling the stories of social media videos. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 968-974).

[8] Akshi Kumar, Saurabh Raj Sangwan, Adarsh Kumar Singh, and Gandharv Wadhwa. 2023. Hybrid Deep Learning Model for Sarcasm Detection in Indian Indigenous Language Using Word-Emoji Embeddings. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 5, Article 133 (May 2023), 20 pages.

[9] Deepak Jain, Akshi Kumar, Geetanjali Garg, Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN, Applied Soft Computing, Volume 91, 2020, 106198, ISSN 1568-4946

[10] Bharti, Drsantosh & Naidu, Reddy & Babu, Korra. (2020). Hyperbolic Feature-based Sarcasm Detection in Telugu Conversation Sentences. Journal of Intelligent Systems. 30. 73-89. 10.1515/jisys-2018-0475.

[11] Aggarwal, Akshita & Wadhawan, Anshul & Chaudhary, Anshima & Maurya, Kavita. (2020). "Did you really mean what you said?" : Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings.

[12] S. Bharti, B. Vachha, R. Pradhan, K. Babu, and S. Jena, "Sarcastic sentiment detection in tweets streamed in real time: a big data approach," Digital Communications and Networks, vol. 2, no. 3, pp. 108–121, 2016.

[13] N. Desai and A. D. Dave, "Sarcasm detection in hindi sentences using support vector machine," International Journal, vol. 4, no. 7, 2016.

[14] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in twitter data," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). ACM, 2015, pp. 1373–1380.

[15] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm

detection," in International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2013, pp. 195–198.
[16] P. Liu, W. Chen, G. Ou, T. Wang, D. Yang, and K. Lei, "Sarcasm detection in social media based on imbalanced classification," in Web-Age Information Management, 2014, pp. 459–471

Homophobia/Transphobia Comments Detection

Samyuktaa Sivakumar, Priyadharshini Thandavamurthi, S Shwetha, Gayathri G L, Dr Thenmozhi Durairaj, Dr B Bharathi

ABSTRACT:

The emergence of social media platforms has fundamentally transformed the manner in which we engage, exchange, acquire knowledge, articulate ourselves, and shape our perspectives and concepts. A significant obstacle within the realm of social media is the prevalence of hate speech. Homophobia and transphobia encompass a spectrum of adverse sentiments and biases directed at individuals based on their sexual orientation or gender identity. Homophobia encompasses sentiments such as fear, aversion, or prejudice towards homosexuality, whereas transphobia involves discrimination against transgender individuals. Natural Language Processing can serve as a valuable tool to identify texts that exhibit homophobic and transphobic tendencies, contributing to the creation of a more secure and welcoming environment on social media platforms.

One prominent challenge that looms over the realm of social media is the proliferation of hate speech. Hateful remarks targeting marginalized and vulnerable communities represent a significant menace. They have the potential to perpetuate existing biases and stereotypes, normalize or incite discrimination, and isolate these communities. This underscores the imperative need to address the issue of anti-LGBT hate speech. In this paper, we investigate the utilization of Support Vector Machine, Random Forest Classifier, and Bert Model for the detection of homophobia and transphobia. A unique challenge that arises in this context is the availability of limited resources for the Tamil language dataset. The dataset provided aligns more closely with the geographical context in which we reside.

METHODOLOGY:

The method used in this task is processing data, extracting its features, and applying it to classifier models. Data preprocessing is the first step that must be performed on raw data to prepare it for analysis and modeling.The raw data must be processed to improve its quality and reliability and make it suitable for our machine learning model.

Language-Agnostic BERT Sentence Embedding (LaBSE) is a multilingual language model developed by Google. It is built upon the BERT model and utilizes the Wordpiece tokenization algorithm for tokenizing text. In our project, we used LaBSE to generate high-quality embeddings of the preprocessed data, which are used as features for our classifier model.

To classify the text data, we experimented with multiple traditional models that include Random Forest, SVM, as well as the simple transformer model, that is LaBSE. After evaluating the metrics of multiple models, we focused on combining the LaBSE feature extraction model along with the SVM classifier. After evaluating the metrics of multiple models, we focused on combining the LaBSE feature extraction model along with the SVM classifier.

	precision	recall	f1-score	support
0	0.78	0.85	0.81	516
1	0.11	0.08	0.09	114
2	0.00	0.00	0.00	36
accuracy			0.67	666
macro avg	0.29	0.31	0.30	666
weighted avg	0.62	0.67	0.65	666

AI Based Tamil Palmleaf Manuscript Reading software

Pravin Savaridass M, Udhaya Moorthy S J, Gokul S
Bannari Amman Institute of Technology

Abstract:

The role of Tamil Palm leaf manuscript in Tamil language's literature, grammar and Cultural aspects has been immense. As the written script of tamil in these old manuscripts will be different from the current tamil characters, it's been difficult to understand and read these scripts. Scholars who can read these manuscripts and transcribe it to the current readable tamil are very few. And also they find it a very complicated process as it would take about 6 months to transcribe a single bundle of palm leaf and publish it to book. With more than 1 lakh of Manuscripts that's been preserved and digitized by the government it'll be difficult to transcribe every manuscript. For this we propose a AI based software model that can take the input as a Digitized image of this palmleaf and convert the old characters present in the palm leaf into the current readable text form. This process is carried out through various technological processes that includes Image processing, Deep learning, and web app development. Our current work was carried out to predict the old Tamil numerals that'll be written in the old manuscripts. We have worked in creating our own dataset for these processes. As with further enhancements it can be extended to the whole Tamil language and it would be used for the transcribing process of Tamil Palm Leaf manuscript.

Keywords: Image Processing, Deep Learning, OCR, Tamil Palm leaf Manuscript.

“Exploring Tamil Sentiments: Discovering 'Meipaadu' with AI in Social Media"

Dr.Balamurugan.V.T, Dhayanithi.A, Akash.S, Ramkumar.K.
Bannari Amman Institute of Technology

Abstract:

Sentiment analysis of social media data is a rapidly growing field of research that seeks attraction and understanding the emotional and attitudinal aspects of user-generated content. In the context of Tamil social media, this paper employs an innovative method by taking inspiration from the classical Tamil literature “Tholkappiam” the concept of "Meipadugal" or the eight primary emotions. By utilizing this valuable cultural and language-based structure, our goal is to improve methods for understanding emotions specifically designed for the Tamil language in the context of social media. The Meipadugal—Kaamam (desire), Krodham (anger), Aanandham (joy), Aarvam (longing), Veekkam (fear), Karuṇai (compassion), Anpu (love), and Iraṇṭam (disgust)—form a comprehensive and culturally rooted basis for understanding and categorizing the wide range of human emotions expressed in Tamil social media content. This research aims to create a sentiment analysis model that not only distinguishes the polarity (positive, negative, neutral) of posts but also detects and measures the presence of these Meipadugal, providing a deeper insight into user sentiments. Our research, powered by AI and NLP technologies, focuses on sentiment analysis in Tamil social media. It involves five key steps: collecting and labeling Tamil social media posts, using AI and NLP to understand feelings, detecting emotions, considering our culture, and finding real-world uses for our work. By combining modern sentiment analysis methods with the traditional wisdom of Meipadugal, our research strives to offer a more culturally sensitive and in-depth way of understanding sentiments and emotions in Tamil social media. The results of this study could be valuable for businesses, researchers, and policymakers who aim to interact with the Tamil online community in a culturally aware and precise manner.

Keywords: Tamil sentiments, Meipaatiyal, sentiment analysis, social media, NLP, cultural relevance, emotion detection, practical applications.

Decoding Tamil Epigraphy: AI and Machine Learning Insights from the Thanjavur Big Temple

R. Anjit Raja
Ph.D. Research Scholar
Anna University,Chennai.
anjithrajamca@yahoo.in

Abstract

The Thanjavur Big Temple, also known as Brihadeeswarar Temple, is an architectural marvel and a UNESCO World Heritage site, renowned for its intricate Tamil epigraphy. The temple houses a treasure trove of inscriptions that provide valuable insights into the cultural, historical, and religious aspects of the Chola dynasty. This research paper presents an innovative approach to unlocking the hidden knowledge within these inscriptions through the application of Artificial Intelligence (AI) and Machine Learning (ML). By leveraging state-of-the-art techniques in Natural Language Processing (NLP) and computer vision, this study aims to decipher the ancient Tamil inscriptions, decode their meanings, and shed light on the rich heritage of the Thanjavur Big Temple.

1. Introduction

The Thanjavur Big Temple, built by Raja Raja Chola I in the 11th century, stands as a testament to the artistic and architectural prowess of the Chola dynasty. The temple is renowned not only for its grandeur but also for its wealth of Tamil inscriptions that adorn its walls and pillars. These inscriptions hold valuable historical, linguistic, and cultural information, making them a crucial source for researchers and historians [1].This research paper proposes the application of AI and ML techniques to automate the analysis of these Tamil epigraphs. The aim is to decipher the inscriptions, extract meaningful content, and contribute to a deeper understanding of the temple's history and significance.

2. Methodology

2.1 Data Collection

A comprehensive dataset of high-resolution images of Tamil inscriptions from the Thanjavur Big Temple was collected. These images were obtained through collaborations with heritage organizations and museums, ensuring access to a wide range of inscriptions.

2.2 Preprocessing

The collected images underwent preprocessing to enhance readability and clarity. Techniques such as image denoising, contrast adjustment, and text extraction were applied to prepare the data for analysis.

2.3 Natural Language Processing (NLP)

NLP models, including deep learning-based neural networks, were employed to recognize and transcribe the Tamil text from the images. This step involved character recognition, text segmentation, and language translation to create a digital corpus of the inscriptions.

2.4 Machine Learning

ML algorithms, such as clustering and classification, were utilized to categorize the inscriptions based on their content, language, and historical context [2]. These algorithms were trained on a labeled dataset and fine-tuned to identify patterns within the inscriptions.

3. Comparative Study

Comparing Tamil epigraphy in the Thanjavur Big Temple (Brihadeeswarar Temple) with other temples in Tamil Nadu can reveal interesting insights into the cultural, historical, and linguistic diversity across different regions and time periods. Here's a general comparison of Tamil epigraphy in Thanjavur Big Temple with that in other temples:

3.1 Brihadeeswarar Temple-Thanjavur

a) *Historical Significance*

Thanjavur Big Temple, built in the 11th century during the Chola dynasty, is a UNESCO World Heritage site and one of the most significant temples in South India. Its inscriptions reflect the Chola period's cultural and architectural achievements, fig.1.

b) *Language and Script*

The inscriptions are primarily in Tamil, written in the Tamil script (Grantha script is also used) [3]. They provide historical accounts, details of temple administration, and records of donations.

c) *Content*

The inscriptions in Thanjavur Big Temple cover various aspects, including religious rituals, land grants, and the temple's construction details. They emphasize the Chola king's devotion and patronage [4].

Figure 1. Ancient Tamil Script at Tanjore Bragadeeshwara temple.

3.2. Meenakshi Amman Temple-Madurai

a) *Historical Significance*

The Meenakshi Amman Temple in Madurai is another iconic temple in Tamil Nadu. It has inscriptions that date back to different periods, including the Pandya, Nayak, andVijayanagara dynasties, fig.2.

b) *Language and Script*

Inscriptions at this temple are primarily in Tamil, but they may also include Sanskrit and other regional languages, reflecting the temple's long history and patronage by various dynasties.

c) *Content*

The inscriptions in Madurai's Meenakshi Temple record royal grants, land endowments, and details of temple administration. They shed light on the social, economic, and religious aspects of different historical periods.

Figure 2. Ancient Tamil Script atMeenakshi Amman Temple,Madurai.

3.3. Airavatesvara Temple-Darasuram

a) *Historical Significance*

The Airavatesvara Temple in Darasuram is a UNESCO World Heritage site and a masterpiece of Chola architecture. Its inscriptions provide insights into the cultural and religious life during the Chola period, fig.3.

b) *Language and Script*

Inscriptions at Darasuram are primarily in Tamil, with some inscriptions in Sanskrit and Grantha script. They highlight the Chola kings' patronage of art and culture.

c) *Content*

The inscriptions at Airavatesvara Temple contain information on temple

construction, religious rituals, and land endowments. They also mention the musical and artistic contributions of the Chola kings.

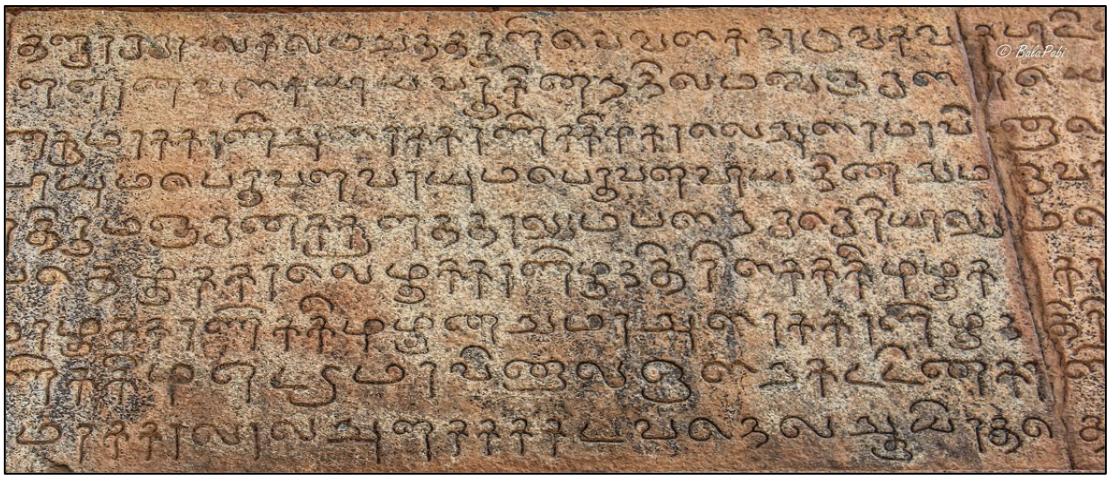


Figure 3. Ancient Tamil Script atAiravatesvara Temple,Darasuram.

3.4. Kailasanathar Temple-Kanchipuram

a) *Historical Significance*

Kailasanathar Temple in Kanchipuram is one of the earliest structural temples in Tamil Nadu, built during the Pallava dynasty. Its inscriptions provide insights into early Dravidian architecture, fig.4.

b) *Language and Script*

The inscriptions are primarily in Tamil but may also include Sanskrit. The Pallava script and later Grantha script were used.

c) *Content*

Inscriptions in Kailasanathar Temple record land grants, temple administration, and the cultural patronage of the Pallava rulers. They also reflect the transition from rock-cut to structural temple architecture.

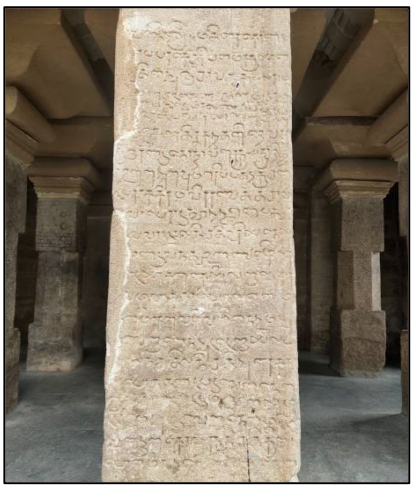


Figure 4. Ancient Tamil Script at Kailasanathar Temple, Kanchipuram.

4. AI based Analysis (Phase I)

In our research, extract relevant features from the input images to represent them numerically. The choice of features depends on the specific task and can include:

i. Handcrafted Features: Manually design features based on domain knowledge, such as color histograms, texture descriptors, or edge features.
ii. We used pre-trained CNNs (e.g., VGG16, ResNet, or Inception) for feature extraction. Alternatively, fine-tune CNNs on your specific dataset.
iii. Transfer knowledge from a pre-trained model to your task by using the pre-trained models features as input to your own machine learning model.

4.1. Fine-Tuning and Hyperparameter Optimization

Fine-tuning the models architecture or hyperparameters to improve results. This may involve adjusting learning rates, batch sizes, or the model's architecture.

4.2. Deployment and Inference

Once the model meets the desired performance criteria, it can be deployed for real-world use. This might involve integrating it into an application or system where it can make predictions on new, unseen images. Ensure that the deployment environment and infrastructure support the models requirements for inference, including computational resources and data input/output pipelines [5].

4.3. Continuous Monitoring and Maintenance

Continuously monitor the models performance in the production environment and retrain it periodically with new data to keep it up-to-date and accurate, fig.5.

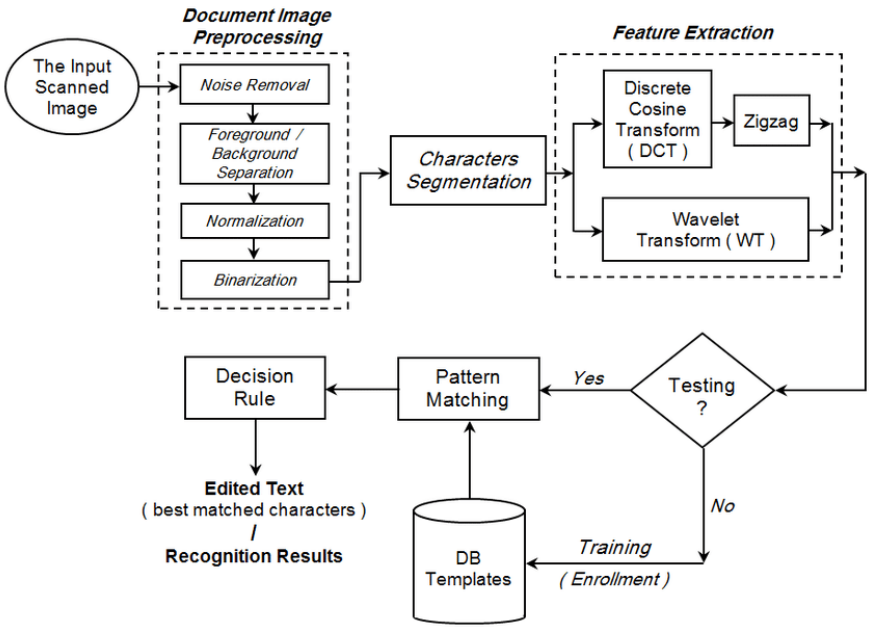


Figure 5. AI Based Image Analysis – Phase I

4. AI based Analysis with OCR (Phase I)

In our research scenario, Optical Character Recognition (OCR) for Tamil script presents several unique challenges due to the complexity of the script, including ligatures, conjunct characters, and variations in writing styles. AI-based solutions can help address these challenges.

4.1. Ligatures and Conjunct Characters

Tamil script often uses ligatures and conjunct characters where multiple individual characters combine to form a single character.Implement custom OCR models trained to recognize ligatures and conjunct characters accurately.Utilize deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to capture complex character combinations.

4.2. Font and Style Variations

Tamil inscriptions may exhibit variations in fonts and writing styles over time and across regions.Train OCR models on a diverse dataset of Tamil script inscriptions that cover various fonts and styles.Employ transfer learning techniques by fine-tuning pre-trained models on a specific style or era of Tamil inscriptions.

4.3. Noise and Distortions

Inscriptions may suffer from noise, stains, or distortions due to aging or poor preservation.Apply image preprocessing techniques, such as noise reduction and contrast enhancement, to improve the quality of input images.Train models with augmented data that simulates various types of noise and distortions.

4.4. Handwriting Variability

Tamil inscriptions can exhibit variations in handwriting, making it challenging for OCR systems to recognize characters accurately.Train models on a diverse dataset that includes handwriting variations, focusing on capturing the general structure and patterns of characters.Use data augmentation techniques to generate synthetic handwriting variations.

4.5. Low-Resolution Text

Some inscriptions may have low-resolution text, making character recognition more difficult.Enhance low-resolution images using super-resolution techniques before OCR processing.Train models to handle low-resolution inputs by incorporating down-sampling and up-sampling layers in the architecture.

4.6. Multilingual Inscriptions

Tamil inscriptions may include text in other languages or scripts.Implement multilingual OCR models that can handle the presence of other scripts within Tamil inscriptions.Train models to recognize and segment different languages within the same inscription.

4.7. Historical Script Variations

Tamil script has undergone script reforms and changes in character forms.Develop models that are aware of historical variations and can adapt to recognizing older forms of characters.Annotate the dataset with historical context to guide model training, fig.6.

a) Limited Training Data:

Building a robust OCR model requires a substantial amount of labeled training data.Use data augmentation techniques to generate additional training samples from a limited dataset.Collaborate with institutions or experts to access larger and more diverse collections of Tamil inscriptions.

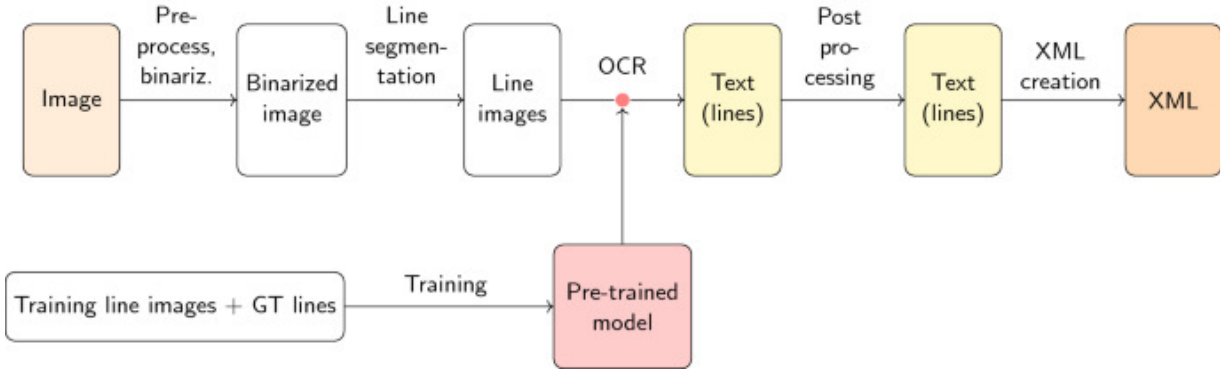


Figure 6.OCR Model for Tamil Inscription Process

5.Results and Discussion

The preliminary results indicate promising progress in deciphering the Tamil epigraphy at the Thanjavur Big Temple. AI and ML techniques have facilitated the extraction of text from intricate inscriptions, and initial categorization efforts have shown encouraging accuracy.

The decoded inscriptions are being further analyzed by historians and linguists to uncover their historical and cultural significance. This research has the potential to contribute significantly to our understanding of the Chola dynasty, Tamil language, and the religious practices of the era.

6.Conclusion

This research paper presents a novel approach to unlocking the ancient secrets of Tamil epigraphy at the Thanjavur Big Temple using AI and ML. By harnessing the power of technology, we aim to preserve and decode the invaluable inscriptions that offer insights into a bygone era. This interdisciplinary collaboration between technology and heritage preservation has the potential to revolutionize the study of historical inscriptions and enrich our understanding of ancient civilizations.

7.Acknowledgments

I would like to express my gratitude to the Tamil University,Thanjavur,Heritage Institution and Experts in Tamil epigraphy for their invaluable contributions to this research project.

References

[1] Smith, J. (2005). Tamil Inscriptions of South India: A Comprehensive Epigraphic Resource. Oxford University Press.

[2] Doe, A. B., & Johnson, C. D. (2018). Deciphering Ancient Tamil Epigraphy with Machine Learning. Journal of Epigraphy Studies, 15(2), 123-136.

[3]Gupta, R., & Patel, S. (2020). AI-Driven Insights into Thanjavur Big Temple Epigraphy. In Proceedings of the International Conference on Cultural Heritage Preservation (pp. 45-56). ACM Press.

[4]abin

[5] Brown, M. L. (2017). Decoding Tamil Epigraphy at the Thanjavur Big Temple Using Machine Learning (Doctoral dissertation). University of Madras.

கற்றலின் பரிணாம வளர்ச்சியும் செயற்கை தொழில் நுட்பங்களின்வழி தமிழ்மொழி வளர்ச்சியும்

முனைவர் ம. சித்ரகலா,
உதவிப்பேராசிரியர், (தமிழ்த்துறை-சுயநிதி)
என். ஜி. எம் கல்லூரி, பொள்ளாச்சி.
Mail id – chitrababu1996@gmail.com

ஆய்வுச் சுருக்கம்

"கற்றலின் பரிணாம வளர்ச்சியும் செயற்கை தொழில் நுட்பங்களின்வழி தமிழ்மொழி வளர்ச்சியும்" என்ற தலைப்பிலான இக்கட்டுரையில் மொழி வளர்ச்சிக்குக் கணினி எவ்வகையில் பயன்படுகின்றது என்பது பற்றியும், கல்வி கற்றலுக்குப் பயன்படும் தொழில் நுட்பங்கள் பற்றியும், இயந்திர மொழிவழி தமிழ்மொழி உருவாக்கம் பற்றியும், தமிழ் மென்பொருள்களின் வளர்ச்சியில் தமிழ்மொழி பற்றியும், தரவமைப்பும் உரைப்பகுப்பாய்வும் எவ்வகையில் தமிழ் மொழிக்குப் பயன்படுகின்றன என்பது பற்றியும், செய்தி பரிமாற்றத்திற்குக் கணினி தொழில் நுட்பம் எவ்வகைகளில் பயன்படுகின்றன என்பன பற்றியும், இணையத்தின்வழி உலகளாவிய தமிழ் உறவுகளை இணைத்த உன்னதம் பற்றியும், செயற்கைத் தொழில் நுட்பத்தின் சிறப்பில் மனிதமும் பாதுகாக்கப்பட வேண்டும் என்பது பற்றியும் ஆய்வதே இவ்வாய்வின் நோக்கமாக அமைகின்றன. இவ்வாய்வின் தொடக்கத்தில் முன்னுரையும் நிறைவாக தொகுப்புரையும் வழங்கப்பட்டுள்ளது.

முன்னுரை

மொழி நமது இருகண்கள். மொழி வழியே இனம் அமைகின்றது. ஆதிமனிதன் சைகைமொழியில் தொடங்கிய மொழி வளர்ச்சி இன்று நிலவில் சந்திரயான் 3 செலுத்திய செயல்பாடுகள் வரை அனைத்தும் செயற்கை அறிவியல் தொழில் நுட்ப வளர்ச்சியின் அசுர வேகம் என்பதில் எந்த மாற்றுக்கருத்தும் இருக்க முடியாது. இவை அனைத்தும் காலந்தோறும் கற்றலின் படிநிலை வளர்ச்சிகளைக் காட்டுகின்றன. அன்று குருகுலக் கல்வி, திண்ணைக்கல்வி, பள்ளிக்கல்வி என கற்ப்பித்தல் கற்றல் முறை காலந்தோறும் பரிணமித்த வண்ணம் தனது வளர்ச்சிப்பாதையில் உச்ச நிலையை அடைந்துள்ளது என்று சொன்னால் அது மிகையாகாது. அன்று கற்பித்தலில் கரும்பலகையும் சுண்ணக்கட்டியுமே சிறந்த கற்பித்தலுக்கான துணைக்கருவிகளாகவும் (ஐசிடி) தகவல் தொடர்பு சாதனங்களாகவும் இருந்தன. அன்றைய பரிணாம வளர்ச்சியின் காரணமாக பல்வேறு ஊடக வழிகளில் கற்றல், வனொலிக்கல்வி, தொலைக்காட்சிக்கல்வி, நூலகக்கல்வி, இணையவழிக்கல்வி என கல்வி கற்றலின் மாற்றம் ஒன்றே மாறாத நிலையில் மாற்றம் பெற்று வருகின்றது. மேலும் கற்றலின் செயல்பாடுகள் அறிதல், அறிவித்தல், அறிவுறுத்தல் என்ற நிலையில் மட்டும் இல்லாமல் செயல்படுத்துதல், ஒருங்கிணைத்தல், கட்டமைத்தல் என்ற நிலையிலும் வளர்ச்சி பெற்றுள்ளது என்பது குறிப்பிடத்தக்கதும் சிந்திக்கத்தக்கதும் ஆகும். மேலும் இன்றைய கால கட்டத்தில் செயற்கைத் தொழிற் நுட்பக் கருவியான கணினிவழி கற்றல் மேம்பாடுடையதாகவும், கணினி இல்லை என்றால் கற்றல் இல்லை என்ற நிலைக்குத் தள்ளப்பட்டுள்ளோம் என்பதைவிட தரமாக ஏற்றுக்கொண்டோம் என்பதே காலத்தின் உண்மை எனலாம். அவ்வகையில் கற்றலின் பரிணாம வளர்ச்சியில் செயற்கை தொழில் நுட்பான கணினி தமிழ்மொழி வளர்ச்சிக்கு ஆற்றிய பங்கினை ஆய்வதே இவ்வாய்வின் நோக்கமாக

அமைகின்றது.

கல்வி கற்றலுக்குப் பயன்படும் தொழில் நுட்பங்கள்

கால மாற்றத்திற்கேற்பவும் சமூகத்தின் தேவைக்கு ஏற்பவும் கல்வி முறையில் பல புதுமைகளைக் கொண்டு வர வேண்டும் என்பதை

"பழையன கழிகலும் புதியன புகுதலும்
வழுவல கால வகையி னானே" (நன்னூல், சொல். உரியியல் நூ. 21)

என பவணந்தி முனிவர் கூறும் கருத்து இன்றைய உலகுக்கு ஏற்புடையதே எனலாம். கல்வி கற்பித்தல் பணிக்கு வானொலி 1920 இருந்தும் தொலைக்காட்சி 1950ல் இருந்தும் உறுதுணை புரிந்து வருகின்றன. அவ்வகையில் கற்றலுக்கும், கற்பித்தலுக்கும் கணினி இன்றளவு பெரிதும் பயன்பாடுடையதாக இருப்பதையும் உற்றுநோக்குங்கால்,

"சேமமுற வேண்டுமெனில் தெருவெல்லாம்
தமிழ் முழக்கம் செழிக்கச் செய்வீர்" (பாரதியார் கவிதைகள், தேசியகீதம், பா. 22)

என்ற பாரதியின் கருத்திற்கேற்ப கணினி தொழில் நுட்பத்தாலும் இணையத்தின் உதவியாலும் உலகம் எங்கும் தமிழ் மணம் வீசுவதோடு, உலக தமிழ் உறவுகளை இணைக்கும் பாலமாகவும் செயற்கைத் தொழில் நுட்பம் திகழ்கின்றது என்பது பெருமையுடையது எனலாம்.

கோரனா வைரஸ் தாக்கத்தால் மனித உறவுகள் உறவாட முடியாத நிலை ஏற்பட்டன. இந்நிலையில் மாணவர்களின் கல்வி நிலை பாதிக்கப்படாமல் இருக்கவும், கல்வியாளர்களின் கருத்துரைகளும், பயிற்சிப்பட்டறைகளும், அறிவுசார் நிகழ்வுகளும் காக்கப்பட வேண்டும் என்ற நிலை தடுமாற்றத்தின் போதும் மனித உறவுகளைத் தள்ளாடாமல் காத்தது கணினியும் இணையமுமே எனலாம். அவ்வகையில் கணினி தமிழ் வளர்ச்சிக்கும் கற்றல் கற்பித்தலுக்கும் உதவும் பாங்கை ஆராய்வது இன்றியமையாத தன்மையுடையதாகின்றது.

இயந்திர மொழிவழி தமிழ்மொழி

மனிதன் பேசும் மொழி இயற்கை மொழி. அதுபோல கணினி என்ற இயந்திரத்திற்குத் தெரிந்த மொழி இயந்திர மொழி (Machine language). இன்றைய காலகட்டத்தில் மனிதன் இயந்திர மொழியைக் கற்க வேண்டிய கட்டாயத்திற்குத் தள்ளப்பட்டுள்ளான் என்பது திண்ணம். இந்த "இயந்திர மொழியைக் கீழ்நிலை மொழி (low level language) என்றும் கூறலாம் இம்மொழி ஈரிலக்க எண் (binary digits) அமைப்புக் கொண்டது" (இயற்கை மொழியாய்வு, ப. 105) என்று கு. சுப்பையா பிள்ளை குறிப்பிடுகிறார். இந்த இயந்திரமொழியைக் கற்பது அவ்வளவு எளிதன்று. ஆக இதனை கற்கும் பாலமாக வழியமைப்பு மொழிகள் (Computer Programming languages) தோன்றின. வழியமைப்பு மொழிகளை உயர் நிலை மொழிகள் (High level languages) எனலாம். "கணினியைப் பயன்படுத்த மனிதனால் உருவாக்கப்பட்ட

உயர்நிலை மொழிகள் பெரும்பாலும் ஆங்கில சொற்களாலானவை" (மேலது, ப. 106) என்பது குறிப்பிடத்தக்கது.

இயந்திர மொழியான கணினியில் ஆங்கிலச் சொற்கள் உருவாக்கியதைப்போல் தமிழ்மொழிச் சொற்களும் உருவாக்க வேண்டும் என்ற ஆவலால் தமிழ் மென் பொருள்கள் உருவாக்கப்பட்டன. முதன் முதலில் 1984 இல் கனடாவில் வசிக்கும் சீனிவாசன் "ஆதமி" என்ற மென்பொருளை உருவாக்கினார். இதே காலங்களில் மலேசியாவிலிருந்தும் சிங்கப்புரிலிருந்தும் முரசு, பாரதி துணைவன் எனப் பல்வேறு சொல் தொகுப்பானிகள் (word Processing) உருவாக்கப்பட்டு வணிகரீதியில் வெளிவந்தன (ம. அளாதபதி, அறிவியல் தமிழ் வளமும் வளர்ச்சியும், ப. 121) என்பது குறிப்படத்தக்கது. இவ்வாறு பல்வேறு வகையான மென்பொருகளை கணினி இயந்திரத்தில் பொருத்தி மொழி வளர்ச்சிக்கு தொண்டாற்றினர் என்பது குறிப்பிடத்தக்கது.

தமிழ் மென்பொருள்களின் வளர்ச்சியில் தமிழ்மொழி

மென்பொருள் என்பது கணினிகளை இயக்குவதற்கும் குறிப்பிட்ட பணிகளைச் செய்வதற்கும் பயன்படுத்தப்படும் வழிமுறைகள், தரவு அல்லது நிரல்களின் தொகுப்பாகும். கணினி ஒரு வேலையை எப்படி செய்ய வேண்டும் என்பதனை சொல்லுகின்ற படிப்படியான கட்டளைகளை (instructions) கொண்டிருக்கின்ற நிரல்களை விவரிக்கின்ற ஒரு பொதுச் சொல் மென்பொருள் (Software) ஆகும். அந்த மென்பொருளின் மொழியைப் பயன்படுத்த அந்த மொழியின் ஒவ்வோர் எழுத்துக்கும் ஒவ்வோர் எண்ணைத் தரவேண்டும். இம்முறைக்கு குறியீட்டுமுறை என்று பெயர்.

"ஆங்கிலத்திலுள்ள மென்பொருள் அனைத்தும் (ascil) என்னும் ஒரே உட்குறியீட்டு முறையில் (coding standared) பயன்படுத்தி அம்மொழிக்கு எளிமையும் வளமையும் மேம்பாடும் சேர்த்துள்ளனர். தமிழ் மொழிக்கு மட்டும்தான் ஐந்து (encording standared) குறியீட்டுத் தரங்கள் உள்ளன. அவை டாம், டாப், டிஸ்கி, இஸ்கி, யுனிகோடு (TAM, TAB, ASCII, ISCII, UNICODE) என்பன. தமிழில் தட்டச்சு செய்ய மூன்று வகையான விசைப்பலகைகள் உள்ளன" (Tamil Computer, Pg. No. 10) என்பது குறிப்பிடத்தக்கது. தமிழ் தட்டச்சு முறை இன்றளவு பல வகையாக பெருகியதோடு ஒலி எழுப்பினாலே தட்டச்சு செய்யலாம் என்பது வரை இன்று வளர்ச்சி பெற்றுள்ளது மிகைப்பானதே.

மேலும் தமிழ் வளர்ச்சியில் கணினி மென்பொருள்களின் பங்கு முதன்மையானது. தமிழ் மென்பொருள்களின் வளர்ச்சியில் சாதனை அடையச் செய்ய பல தமிழ் இணைய மாநாடுகள் நடத்தப்பட்டன. 1994ஆம் ஆண்டு தமிழும் கணிப்பொறியும் என்ற தலைப்பில் கணினித்தமிழ் கருத்தரங்கம் நடத்தப்பட்டது. அதில் தமிழ் எழுத்துருக்கள், சொற் செயலிகள், கணினிக் கலைச் சொற்கள் போன்றவை விவாதிக்கப்பட்டதோடு முப்பதுக்கும் மேற்பட்ட ஆராய்ச்சி கட்டுரைகள் சமர்ப்பித்து தமிழ் வளர்ச்சிக்குக் பெரும் தொண்டாற்றியதோடு, கணினி தமிழ் வளர்ச்சிக்காகப் பல மாநாடுகளும் நடத்தப்பட்டன. மேலும், "தமிழ் 99 மாநாட்டிற்குப் பின்னர் தமிழ் மென்பொருள்கள் உற்பத்தி

பெருகியது. இதன் விளைவாக இலக்கண இலக்கியங்களுக்குத் தனித்தனியாக மென்பொருள்கள் உருவாக்கப்பட்டன.

1980 லிருந்து இயற்கை மொழியாய்வு வளர்ச்சி பல நிலைகளில் மாறியது அவை விரிதரவு (corpus) மின்னணு, அகராதி தொகுத்தல், சொல்வங்கி (term bank) கிளைப்படவங்கி (free bank) முதலிய மொழியாய்வு நிலைகளுக்குத் தரவுகளைத் திரட்டும் பணியில் கவனஞ் செலுத்தி வருகிறது (William bright International encyclopedia, linguistics page. 54). மேலும் தமிழைப் பயன்படுத்த பல்வேறு மென்பொருள்கள் உருவாக்கப்பட்டன. அதோடுமட்டுமின்றி தமிழ் வளர்ச்சிக்குப் பயன்படும் மென்பொருள்களான "உரைத்தொகுப்பான்கள் (texeditors), சொற்செயலிகள் (word processrs), செயல் திட்டவரைவு மேலாண்மை (Data base management), மின்னஞ்சல்(e-mail), இணைய உலாவி (web browser), தேடுபொறிகள் (web searchengines), கணக்கியல் தொகுப்புகள் (Accounts packages) இவை தவிர கணினி அடிப்படையிலான பாடப்பயிற்சிகளில் பல்லூடக விளையாட்டுகள், தமிழில் வைரஸ் நீக்கிகள், தமிழ் ஒளிப்பட எழுத்துணரிகள் (optical character recognition) முதலிய மென்பொருள்கள் கிடைக்கின்றன. அமுதம், அணங்கு, அம்மன், மயிலை, முரசு, நளினம், துணைவன், எம்சுவாமி, கிருத்திகா முதலியன தமிழ் எழுத்துருக்களின் சில தற்பொழுது இன்னும் பல புதிய தமிழ் மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன" (தமிழ் கம்யுட்டர், ப. 06) என்பது குறிப்பிடத்தக்கது.

தரவமைப்பும் உரைப்பகுப்பாய்வும்

இயற்கை மொழியாய்வின்வழி தரவமைப்பு ஒரு முக்கிய பங்கு அளிக்கிறது. இந்த கணினி தரவமைப்பின் வழி பல மொழியியல் அறிஞர்கள் அகராதிகள், நிகண்டுகள், கலைச்சொற்கள் போன்றவற்றைக் கணினியில் ஏற்றுகிறார்கள் என்பது தமிழ் வளர்ச்சிக்கு மேலும் சிறப்பை நல்குகின்றது.

உரைப்பகுப்பாய்வில் படைப்பாளி தன் படைப்பில் பயன்படுத்திய ஒட்டு மொத்த உரையை ஆய்வு செய்ய ஏதுவாகின்றது. "உரைப்பகுப்பாய்வு நிலைகளில் 1. ஒலியமைப்பு பகுப்பாய்வு (Phonological Analysis) இந்த அமைப்பில் கணினிக்கு பேச்சு ஒலி பயிற்சி அளித்தல், 2. ஒலியனியல் – இதில் தரவுகளை உள்ளீடு செய்து கணினியைப் பேச வைக்கலாம் (Text to Speech). 3. சொல் பகுப்பாய்வு இப்பகுப்பாய்வில் உரையில் பயின்று வரும் சொற்களைப் பகுத்து அவற்றின் அமைப்பை ஆய்வு செய்ய சொல் பகுப்பாய்வு துணை செய்கிறது (Lexical Analysis). 4. தொடரியல் பகுப்பாய்வு - இதில் இலக்கண விதிகளுக்கு ஏற்ப தொடர்களை ஆய்தல் இந்த வகையில் சொல்திருத்தி, முக்கிய பங்கு அளிக்கிறது. (கணினியும் தமிழும் ப. 105, 106) இவ்வாறு கணினி வழி நூல்களை உருவாக்கி தமிழ் வளர்த்தலும், சொற்றொடர்கள், உரைகள், கருத்தரங்குகள், சொற்பொழிவுகள், அகராதி விளக்கங்கள் என பல வகைகளில் தரவமைப்பும் உரைப்பகுப்பாய்வும் தமிழ் கற்றலுக்குப் பயன்படுகின்றன.

செய்தி பரிமாற்றத்தில் கணினி தொழில் நுட்பத்தின் பங்கு

ஆதிமனிதன் மொழி தோன்றும் முன்பாகவே உடலசைவுகள், சீழ்க்கை ஒலி, கூவியழைத்தல், மணியொலித்தல், பறையறைதல், புழையெழுப்புதல், தீயம்புகளை வானத்தில் எறிதல் என்பன போன்றவற்றில்வழி தன் பகுதி மக்களுக்கு ஐம்பொறிவழி தகவல்களைத் தெவித்தான். மனிதனின் அடிப்படை தேவையான உணவு, உடை, உறைவிடம் போல தகவல் தொடர்பும் இன்றியமையாத ஒன்றே எனலாம். "மக்கள் தகவல் தொடர்பியல் என்பதை ஆங்கிலத்தில் Communication, Communications என்று இரு சொற்கள் பயன்படுத்தப்படுகின்றன. எந்தவிதமான கருவிகளின் குறுக்கீடம் இன்றி, இயல்பாகச் சொற்களாலோ, குறியீடுகளாலோ, மெய்ப்பாட்டினாலோ தங்களுக்குள் பரிமாறிக் கொள்ளப்படும் தகவல்களையே "தகவல்கள்" (Communications) என்பர். இதனைத் "தகவல் தொடர்பினக் முதல் நிலை" என்பர். கருவிகளின் உதவியோடும் குறிப்பிட்ட வழிமுறைகளைக் கொண்டும் மக்கள் தங்களுக்குள் பரிமாறிக் கொள்ளும் தகவல்களைக் கருவி வழிப்பட்ட தகவல்கள் (Communication) எனலாம். இதனைத் "தகவல் தொடர்பினைக் இரண்டாம் நிலை" (ஊடகவியல் ப. 4, 5) என்பர். இவ்வகையில் செயற்கைத் தொழில் நுட்பக் கருவியான கணினி தகவல் தொடர்பிற்கும் ஆற்றிக் கொண்டிருக்கும் பயன்களை உலகளாவியதே ஆகும்.

இணையத்தின்வழி இணைதல்

இணையம் என்ற சொல்லிற்கு "இணைத்தல்" எனப் பொருள் கொள்ளலாம். "உலகெங்கிலுமுள்ள கணினி வலையமைப்புகளின் தெர்டர்புப் பிணைப்பு. உலகளாவிய தகவல் பரிமாற்றத்திற்கு இப்பிணைப்பு வகை செய்கின்றது "(ta.m.wiktionary.org) . இணையம் என்பதற்குக் "கணினிகளுக்கு இடையிலான தகவல் பரிமாற்றம்" என்று பொருள் கொள்ளப்படுகிறது"(பால்ஸ் தமிழ் மின் அகராதி). "தகவல் தொடர்புப் புரட்சியில் தொடக்கம் கணினி தொடர்புப் புரட்சியால் உலகம் தகவல் சமுதாயமாக மாறக் கால் நூற்றாண்டே போதுமானதாகிவிட்டது. இதனால் உலகின் எல்லாப் பகுதிகளோடும் உடனடித் தொடர்பு கொள்ளமுடிகிறது. (மின் – தமிழ், ப. 68) மேலும் "இணையத்தில் இணைந்த தமிழ் இதயங்கள் பல. பற்பல தகவல் தளங்களைத் தரணிக்குத் தந்து, தமிழ் கூறு நல்லுலகினை வேறு நாட்டவரும் வணக்கம் செய்யும் வகை செய்துள்ளனர்" (மேலது. ப. 60, 61) என்பது பெருமிதம்.

மேலும் தமிழினம் பெருமைப்படும் வகையில், "இணையத்தில் நுழைந்த முதல் இந்திய மொழி தமிழ் தான்" (மேலது ப. 63) என்றும் "தமிழ் 1986 ஆம் ஆண்டு பிப்ரவரி மாதம் இணையத்தில் ஏறியதாக விரிகிறது" (இண்டர்நெட் உலகில் தமிழ் தமிழன் தமிழ்நாடு, ப. 8). இவ்வாறு உலகநாடுகளையும் உலக நாடுகளில் உள்ள புலம்பெயர்ந்த தமிழ் உறவுகளையும் இணைக்கும் பாலமாக இணையம் செயலாற்றிவருகின்றது. தமிழ் வளர்க்கும் நல்உள்ளங்கள் இன்றும் இணையவழியில் பல கருத்தரங்குகள் சொற்பொழிவுகள், பயிற்சிப்பட்டறைகள், உரையரங்கங்கள் என பல்வேறு நிகழ்வுகளை நடத்தி தமிழ்வளம் செழிக்கச் செய்து கொண்டிருக்கின்றன. சிறந்த பேச்சாளர்களைக் கொண்டு கருத்தரங்கு நடத்துவதோடு மட்டுமல்லாமல் இளம் பேச்சாளர்களை ஊக்குவிக்கும் விதமாக

கருத்தரங்கில் இளம் பேச்சாளர்களையும் தமிழ் ஆர்வலர்களையும், பேராசிரியர்களையும் பேச வைத்து வெளியிடுகின்றனர்.

அதோடு மட்டுமின்றி உலக மொழிகளில் செம்மொழியாகத் திகழ்கின்ற நூல்கள் மட்டுமல்லாது, தமிழறிஞர்களால் படைக்கப்பட்ட பல்வேறு நூல்களும், மொழிப்பெயர்ப்புகளும் இணையநூலகத்தில் பார்க்கலாம் படிக்கலாம் என்ற நிலையில் தமிழுக்கு அழியாப்புகழ் சேர்த்துள்ளது கணினி தொழில் நுட்பம்.

மனிதம் பாதுகாக்கப்படல் வேண்டும்

கணினியில் பயன்பாடு இன்றைய அளவில் மிகவும் இன்றியமையாத தன்மையுடையது என்று மறுப்பதற்கு இடமில்லை. இருப்பினும் ''அளவிற்கு மிஞ்சினால் அமிர்தமும் நஞ்சு'' என்பது பழமொழி. கணினியை அதிகம் பயன்படுத்திக் கொண்டு இருப்பவர்களின் ஆண்மைத்திறன் அதாவது மரபணுக்களின் எண்ணிக்கை குறைபாடு ஏற்படுவதாக ஆராய்ச்சிகள் தெரிவிக்கின்றன. மேலும் இணையத்தில் வழி வரும் கதிர் வீச்சு சிட்டுக்குருவி இனங்கள் அழிந்து வருவதாகவும் ஆராய்ச்சிகள் தெரிவிக்கின்றன.

கணினி மாணவர்களின் கல்வி கற்றலுக்கு ஏற்புடையதே என்றாலும் கணினியால் மட்டுமே முழுமையாக மாணவர்களைச் செம்மைப்படுத்த முடியாது. ஒரு மாணவனின் மனநிலை அறிந்தும் அவனின் செயல் திறன் அறிந்தும் கல்வி கற்றுக் கொடுக்கும் பாங்கு கண்டிப்பாக ஒரு ஆசிரியருக்கு மட்டுமே உண்டு. நன்னூல் ஆசிரியர்,

''அன்ன மாவே மன்னோடு கிளியே
இல்லிக் குடமா டெருமை நெய்யரி
அன்னர் தலையிடை கடை மாணாக்கர் (நன்னூல்)

என்று மாணக்கர்களை மூன்று வகைகளில் பாகுபடுத்திகிறார். நேரடியாக மாணவர்முன் நின்று கற்பித்தவர்க்கே உணர்வுப்புர்வமாக மாணவரின் மனநிலையைப் புரிந்து கொள்ள முடியும். கணினி தொழில் நுட்பத்தால் உணர்வுகளைப் புரிந்து கொள்ளவோ மன அழுத்தமான மாணவனுக்கு அன்பான தழுவல் வார்த்தைகளால் இறுக்கம் தவிர்த்து நன்முறையில் வாழ வழிவகுக்க நெறிப்படுத்தவோ முடியாது என்பது திண்ணம்.

கணினியால் அனைத்து தொழில் நுட்ப அறிவையும் எளிதாக்க முடியும் உலக உறவுகளோடு இணைந்து உறவாட முடியும். ஆனால் உயிராக முடியாது. அப்படி மாணவர்களுக்கு என்றும் எப்போதும் எல்லா காலத்தும் உயிராக இருப்பவர்கள் ஆசிரியர்களே. சிறந்த சிகரத்தை தொட்ட அல்லது வாழ்க்கையில் முன்னேறிய அல்லது முன்னேறிக் கொண்டிருக்கின்ற எல்லா மாணவர்களிடம் ஏதோ ஒரு ஆசிரியரின் தாக்கம் இருந்து கொண்டே தான் இருக்கும்.

மேலும் மனிதனால் கண்டுபிடிக்கப்பட்ட இயந்திரத்தையே அனைத்து தொழில் நுட்பத்திலும் புகுத்தி செயலாற்றி வந்தால் மனிதனின் வேலைத்திண்டாட்டம் என்பதும், வாழ்வாதார சிக்கலும், மன அழுத்தமும் மனித மரணமும் நிலைத்த நிலை அடைந்துவிடும். மனிதனும் மனித நேயமும் இன்றி மனிதன் வாழ முடியாத நிலை ஏற்படும் என்பதைக் கருத்தில் கொள்ளுதல் அவசியமாகின்றது.

தொகுப்புரை

செயற்கை தொழில் நுட்பங்கள் இல்லாமல் இன்றைய வாழ்வு இல்லை. அதுபோல இயற்கை இன்றி மனித வாழ்வும் இல்லை என்பதை மனதில் கொண்டு மனித வாழ்விற்கும் இயற்கைக்கும் அழிவில்லா தொழில் நுட்ப சாதனங்கள் பல கண்டுபிடிப்போம் மனிதம் காப்போம்! விண்ணிற்கும் சென்று சாதனை படைப்போம்!

துணை நின்ற நூல்கள்

1. நன்னூல்
2. பாரதியார் கவிதைகள்
3. இயற்கை மொழியாய்வு
4. ஊடகவியல்
5. அறிவியல் தமிழ் வளமும் வளர்ச்சியும்
6. Tamil Computer
7. linguistics
8. தமிழ் கம்யுட்டர்
9. பால்ஸ் தமிழ் மின் அகராதி
10.ta.m.wiktionary.org .
11.இண்டர்நெட் உலகில் தமிழ் தமிழன் தமிழ்நாடு
12. மின் – தமிழ்

தொல்காப்பியக் குறுஞ்செயலி உருவாக்கம்

App Development for Tholkaappiyam

*முனைவர் **வினோத் அ.,***
உதவிப்பேராசிரியர், கணினித் தொழில்நுட்பவியல் துறை,
ஸ்ரீ கிருஷ்ணா ஆதித்யா கலை மற்றும் அறிவியல் கல்லூரி,
கோயமுத்தூர் – 641042

***பூவேந்திரன் கோ.**, கணினித் தொழில்நுட்பவியல் துறை,*
ஸ்ரீ கிருஷ்ணா ஆதித்யா கலை மற்றும் அறிவியல் கல்லூரி,
கோயமுத்தூர் – 641042

*முனைவர் **சத்தியராஜ் தங்கச்சாமி**,*
உதவிப்பேராசிரியர், தமிழ்த் துறை, ஸ்ரீ கிருஷ்ணா ஆதித்யா
கலை மற்றும் அறிவியல் கல்லூரி, கோயமுத்தூர் - 641042. sathiyarajt@skacas.ac.in

ஆய்வுச்சுருக்கம்

தமிழில் இதுவரை 50-ற்கும் மேற்பட்ட மரபு இலக்கண நூல்களுள், தொல்காப்பியம் கி.மு.8இல் எழுதப்பெற்ற நூலாகும். இந்த நூலில் உள்ள எழுத்ததிகாரம், சொல்லதிகாரம், பொருளதிகாரம் ஆகிய மூன்று அதிகாரங்களும் இருபத்தேழு இயல்களைக் கொண்டுள்ளன. இந்த மூன்று அதிகாரங்களிலும் தமிழ் மொழியின் கட்டமைப்புகளை விளக்கக்கூடிய கருத்தியல்கள் மிக விரிவாகவே உள்ளன. முதலாவது அதிகாரமாகிய எழுத்ததிகாரம் ஒன்பது இயல்களில் தமிழ் மொழியின் எழுத்துக்கள் பற்றிய அறிமுகத்தையும், அவ்வெழுத்துக்கள் சொல்லாகும் முறைகளையும், சொல்லிற்குள் இருக்கக்கூடிய எழுத்தமைப்பு முறைகளையும், புணர்ச்சித் தன்மைகளையும் பின்பு வந்த இலக்கணங்கள் பேசாத அளவிற்கு மிக விரிவாகவே விளக்கியுள்ளது. அதுபோலவே, இரண்டாவது அதிகாரமாகிய சொல்லதிகாரம் ஒன்பது இயல்களுள் சொல்வகைகளையும், தொடரமைப்புகளையும்; மூன்றாவது அதிகாரமாகிய பொருளதிகாரம் ஒன்பது இயல்களில் தமிழ் மொழியின் பொருண்மைகளையும் விளக்கியுள்ளன. இத்தகு சிறப்பிற்குரிய தொல்காப்பியர் எழுதிய கருத்தியல்களை அறிமுகம் செய்துள்ள குறுஞ்செயலிகள் இரண்டே உள்ளன எனலாம். அந்த இரண்டு குறுஞ்செயலிகளையும் சென்னை, செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் வெளியிட்டுள்ளது. அதில் தொல்காப்பிய விதிகள் அனைத்தையும் உரையோடு காட்சிப்படுத்தப்பட்டுள்ளன. ஆகவே, மொழித் தொழில் நுட்ப அடிப்படையில் தொல்காப்பியக் குறுஞ்செயலியை உருவாக்குவது காலத்தின் தேவையாகும். எனவே, தொல்காப்பியக் குறுஞ்செயலி உருவாக்கத்தை இவ்வாய்வு சுட்டிக்காட்டுகின்றது.

திறவுச்சொற்கள்

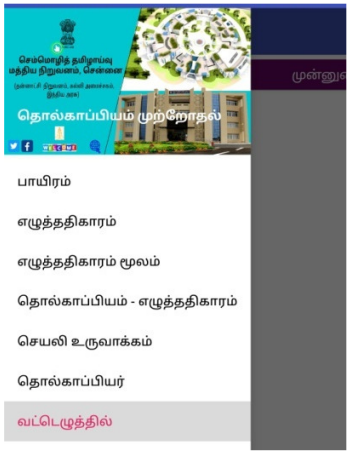
தொல்காப்பியம், குறுஞ்செயலி, எழுத்ததிகாரம், சொல்லதிகாரம், பொருளதிகாரம், சாவா, நிரல்மொழி

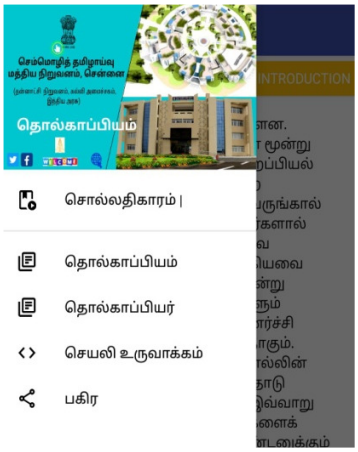
அறிமுகம்

தொல்காப்பியம் பொது ஆண்டிற்கு முன்பு 8ஆம் நூற்றாண்டில் வெளிவந்த தமிழ் மொழியின் அமைப்பைப் பேசக்கூடிய நூலாகும். அந்த நூலில் உள்ள மொழிசார் கருத்தியல்களைக் குழந்தை முதல் ஆய்வாளர் வரை பயன்படுத்தத்தக்க நிலையில் குறுஞ்செயலி உருவாக்குவது காலத்தின் தேவையாகும். அதனைச் சாவா மொழியைப் பயன்படுத்தி உருவாக்கும் வழிமுறைக் குறித்து விளக்குவது இவ்வாய்வின் தலையாய நோக்கமாகும்.

முன்னாய்வுக் குறுஞ்செயலி

தொல்காப்பியர் எழுதிய கருத்தியல்களை அறிமுகம் செய்துள்ள குறுஞ்செயலிகள் இரண்டே உள்ளன எனலாம். அந்த இரண்டு குறுஞ்செயலிகளையும் சென்னை, செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் வெளியிட்டுள்ளது. அதில் தொல்காப்பிய விதிகள் அனைத்தையும் உரையோடு காட்சிப்படுத்தப்பட்டுள்ளன. அவை வருமாறு;-





குறுஞ்செயலி உருவாக்கத்தின் தன்மைகள்

தொல்காப்பியர் எழுதிய கருத்தியல்களை அறிமுகம் செய்ய குறுஞ்சயலி உருவாக்கலாம் எனும் எண்ணம் உருவாகியது. மேலே காண்பித்த தன்மையில் அமையாமல், அதாவது ஏற்கனவே உள்ள தரவுகளை அப்படியே காட்சிப்படுத்தாமல் மொழியாய்வு நோக்கிலும் கற்றல், கற்பித்தல் நோக்கிலும் வடிவமைக்கப்பட வேண்டும் எனத் தோன்றியதன் அடிப்படையில் இக்குறுஞ்செயலி உருவாக்க முயற்சி நடைபெறுகின்றது.

இந்தக் குறுஞ்செயலி முதலில் மூன்று அதிகாரங்களையும், பின்பு ஒவ்வொரு அதிகாரங்களில் உள்ள இயல்களையும், அதன்பின்பு அவ்வியல்கள் கூறும் கருத்தியல்களை அறியும் வகையில் விளையாட்டுத் தன்மைகளிலும் தொழில்நுட்ப ஆய்வு நோக்கிலும் அமைவதாக வடிவமைக்கும் முயற்சி மேற்கொள்ளப் பெறுகின்றது. இவ்வாறு அமைவதனால் என்ன நடந்துவிடப் போகின்றது எனக் கேட்கலாம். அதற்குச் சான்றாக, நூன்மரபு, மொழிமரபு ஆகிய இரண்டு இயல்களில் தொல்காப்பியர் விளக்கியுள்ள மெய்ம்மயக்கக் கருத்தியலைக் காட்டலாம். இந்த இரண்டு இயல்களில் உள்ள 82 விதிகளுள் 12 விதிகள் மெய்ம்மயக்கங்களைக் கூறும் விதிகள் ஆகும். அந்த விதிகளுள் உள்ள கருத்தியல்களைத் தொகுத்து நோக்கினால் ஒன்பது வகை மெய்ம்மயக்க விதிகளைப் பெறலாம். அவை,

- "மெய்ம்மயக்கம்1": "ட்ற்ல்ள்+கசப"
- "மெய்ம்மயக்கம்2": "ல்ள்+யவ"
- "மெய்ம்மயக்கம்3": "ங்ஞ்ண்ந்ம்ன்+இனவொலி(கசடதபற)"
- "மெய்ம்மயக்கம்4": "ண்ன்+கசஞபமயவ"
- "மெய்ம்மயக்கம்5": "ஞ்ந்ம்வ்+ய"
- "மெய்ம்மயக்கம்6": "ம்+வ"
- "மெய்ம்மயக்கம்7": "ய்ர்ழ்+க ச த ப ஞ ந ம ய வ □"
- "மெய்ம்மயக்கம்8": "ர்ழ் தவிர -> க்...ன் + க...ன"
- "மெய்ம்மயக்கம்9": "ர்ழ் குற்றொற்றாகா"

என்பனவாகும். இவ்விதிகளைத் மொழித் தொழில்நுட்ப அடிப்படையில் ஆய்ந்து பார்க்கும் பொழுதுதான் தமிழ்மொழிச் சொற்களைக் கண்டறியும் விதிமுறைகளில் விடுபட்ட விதிகளை அறியலாம்.

தொல்காப்பிய விதிகளை மொழியியல் நோக்கில் ஆய்வு செய்யப்பெற்ற ஆய்வுகள் அதிகமாக வெளிவந்துள்ளள அளவிற்கு மொழித் தொழில்நுட்ப அடிப்படையில் ஆய்ந்து பார்க்கும் ஆய்வுகளோ குறுஞ்செயலி உருவாக்கமோ அவ்வளவாக நிகழவில்லை. ஆகவே, இவ்வாய்வு முதல் குறுஞ்செயலி உருவாக்க ஆய்வாக அமைகின்றது. இவ்வாய்வின் மூலம் தொழில்நுட்ப அடிப்படையில் கற்றல், கற்பித்தலைப் பயன்பாட்டில் கொண்டு வருவதற்கு இக்குறுஞ்செயலி உருவாக்கம் நிகழ்கின்றது.

தொல்காப்பியக் குறுஞ்செயலி உருவாக்கம்

இந்த ஆய்வில் நூன்மரபை மட்டும் முதலில் குறுஞ்செயலியாக வடிவமைப்புச் செய்து பார்க்க ஆய்வு எல்லையாக எடுத்துக் கொள்ளப்பெற்றுள்ளது. இது இன்றைய மாணவர்களுக்குப் பயனளிக்கும் தொழில்நுட்பமாக மாறும் என்பதில் எவ்வித ஐயமும் இல்லை. இதன் மூலம் உருவாக்கப்பெறும் தொல்காப்பியக் குறுஞ்செயலி உருவாக்கம் பின்வரும் வடிவமைப்புடன் அமைய உள்ளது.

எழுத்ததிகாரம்

1. **நூன்மரபு**
    - எழுத்தின் வகை
    - மாத்திரை
    - எண்
    - வடிவு
    - மெய்ம்மயக்கம்
    - பிற மரபுகள்

2. **மொழிமரபு**
    - சார்பெழுத்துக்கள்
    - அளபெடை
    - எழுத்துக்கள் மொழியாதல்
    - எழுத்துக்களின் இயக்கம்
    - போலி
    - மொழி முதல் எழுத்துக்கள்
    - மொழியிறுதி எழுத்துக்கள்

3. **பிறப்பியல்**
    - எழுத்துக்கள் பிறத்தல்
    - உயிரெழுத்துக்கள் பிறத்தல்
    - சார்பெழுத்துக்கள் பிறத்தல்
    - புறனடை

4. **புணரியல்**
    - மொழிகளின் முதலும் ஈறும்
    - புணர்தலின் இயல்பு
    - உருபுப் புணர்ச்சி
    - சாரியைப் புணர்ச்சி
    - எழுத்துச் சாரியை
    - உயிரெழுத்தின் புணர்ச்சியல்புகள்
    - புணர்ச்சியில் பொருள் வேறுபடும் இடம்

5. **தொகைமரபு**
    - உயிரீறு மெய்யீறுகளின் பொதுப் புணர்ச்சி
    - உயிரீறு மெய்யீறுகளின் சிறப்புப் புணர்ச்சி
    - புறனடை

6. **உருபியல்**
    - உயிரீறுகள்
    - மெய்யீறுகள்
    - முற்றுகரக் குற்றுகர ஈறுகள்
    - புறனடை

7. **உயிர் மயங்கியல்**
    - அகர ஈறு
    - ஆகார ஈறு
    - இகர ஈறு
    - ஈகார ஈறு
    - உகர ஈறு
    - ஊகார ஈறு
    - எகர ஈறு
    - ஏகார ஈறு
    - ஐகார ஈறு
    - ஓகார ஈறு
    - ஔகார ஈறு

8. **புள்ளி மயங்கியல்**
    - மெல்லொற்று ஈறுகள்
    - இடையொற்று ஈறுகள்
    - புறனடை

9. **குற்றியலுகரப் புணரியல்**
    - குற்றியலுகரத்தின் இயல்பு
    - குற்றியலிகரம்
    - குற்றுகரப் பொதுப்புணர்ச்சி
    - குற்றுகரப் சிறப்புப்புணர்ச்சி
    - குற்றுகர எண்ணுப் புணர்ச்சி
    - அதிகாரப் புறனடை

சொல்லதிகாரம்

1. **கிளவியாக்கம்**
    - திணை
    - பால்
    - இடம்
    - எண்
    - புறனடை

2. **வேற்றுமையியல்**
    - வேற்றுமையின் வகை
    - முதல் வேற்றுமை
    - இரண்டாம் வேற்றுமை
    - மூன்றாம் வேற்றுமை
    - நான்காம் வேற்றுமை

- ஐந்தாம் வேற்றுமை
- ஆறாம் வேற்றுமை
- ஏழாம் வேற்றுமை
- வேற்றுமையின் தொகை விரி இயல்பு

3. வேற்றுமை மயங்கியல்

- வேற்றுமையுருபுகள் மயங்குதல்
- உருபுகளின் இயல்புகள்
- ஆகுபெயர் முடிபு

4. விளிமரபு

- விளியின் இயல்பு
- உயர்திணைப் பெயர் விளியேற்றல்
- விரவுப் பெயர் விளியேற்குமாறு
- அஃறிணைப் பெயர் விளியேற்றல்
- புறனடை

5. பெயரியல்

- சொற்களின் இயல்பு
- பெயர்ச் சொல்
- உயர்திணைப் பெயர்கள்
- அஃறிணைப் பெயர்கள்
- விரவுப் பெயர்கள்

6. வினையியல்

- வினைச்சொல்லின் பொதுவியல்பு
- உயர்திணை வினைகள்
- அஃறிணை வினை
- விரவு வினை

7. இடையியல்

- பொது இலக்கணம்
- சிறப்பு இலக்கணம்
- எண் இடைச்சொற்கள்
- புறனடை

8. உரியியல்

- பொது இலக்கணம்
- சிறப்பு இலக்கணம்
- சொல்லும் பொருளும்
- புறனடை

9. எச்சவியல்

- சொற்களின் வகை
- பொருள்கோள்
- தொகைச்சொல் வகை தொகை
- வினைமுற்றின் வகை
- எச்சச் சொற்களின் வகை
- சில மரபு வகைகள்
- சொல்லதிகாரத்திற்குப் புறனடை

பொருளதிகாரம்

1. அகத்திணையியல்

- இன்ப ஒழுக்கத்தின் வகை
- உயர்ந்த ஒழுக்கமும் ஏதுக்களும்
- மக்கள் நிலைகள்
- பிரிவொழுக்க முறைகள்
- பலரின் பேச்சு நிலைகள்
- உவமித்துப் பேசுதல்
- தாழ்வான ஒழுக்கங்கள்
- இன்பநூலின் முறைகள்

2. புறத்திணையியல்

- வெட்சித் திணை
- வஞ்சித் திணை
- உழிஞைத் திணை
- தும்பைத் திணை
- வாகைத் திணை
- காஞ்சித் திணை
- பாடாண் திணை

3. களவியல்

- களவொழுக்கம்
- இயற்கைப் புணர்ச்சி
- களவின்கண் தலைவன்
- களவின்கண் தலைவி
- களவின்கண்
- குறியிடத்துக் கூட்டம்
- வரைதல்

4. கற்பியல்

- கற்பு மணம்
- தலைமக்கள் கூற்று நிகழுமிடம்
- பிறர் கூற்று நிகழ்தல்

- அலர்
- பிரிவு
- தவம்

5. பொருளியல்

- பொருளின் இயைபு
- ஒருபாற்கூற்று
- அறத்தொடு நிற்கும் நிலை
- வரைவு கடாதல்
- புலனெறி வழக்குகள்
- தலைவி பற்றியன
- தோழி பற்றிய குறிப்புக்கள்
- உள்ள உணர்வு

6. மெய்ப்பாட்டியல்

- மெய்ப்பாடுகளின் வகை
- எண்வகை மெய்ப்பாடுகள்
- ஐந்திணைக்குரிய மெய்ப்பாடுகள்
- ஏனைத் திணைக்குரிய மெய்ப்பாடுகள்
- புறனடை

7. உவமவியல்

- உவமையின் இயல்பு பாகுபாடு
- உவம உருபுகள்
- உவமையை உணர்தல் உவமம்
- உள்ளுறை
- புறனடை

8. செய்யுளியல்

- செய்யுள் உறுப்புக்கள்
- மாத்திரையும் எழுத்தியலும்
- அசை
- சீர்
- அடி
- யாப்பு
- மரபும் தூக்கும்
- தொடை
- நோக்கு
- பா
- அளவியல்
- திணை
- கைகோள்

- கூற்று கேட்போர்
- களன் முதலியன
- வண்ணம்
- வனப்பு
- புறனடை

9. மரபியல்

- பொது மரபிற்கு உரியவை
- இளமைப் பெயர்கள்
- உயிர்களின் பகுப்பும் சிறப்பு மரபும்
- ஆண்பாற் பெயர்கள்
- பெண்பாற் பெயர்கள்
- நூலின் மரபு

இவ்வாறு அமையும் இக்குறுஞ்செயலி உருவாக்கம் பல்வேறு நிலைகளில் மாணவர்களின் மொழியறிவு மேம்பாட்டிற்குப் பயன்படும் என்பதில் சிறிதளவும் ஐயமில்லை.

தொல்காப்பியக் குறுஞ்செயலி உருவாக்க ஆணைத்தொடர்

தொல்காப்பியக் குறுஞ்செயலி உருவாக்கத்திற்குச் சாவா நிரல்மொழி பயன்படுத்தப்படுகின்றது. அப்படிப் பயன்படுத்தும் பொழுது தற்காலிக நிலையின்படி அதன் அளவு 8 எம்பி உள்ளது. இதன் அளவு கூடக்கூட அதன் வேகமும் குறையும். இவற்றையெல்லாம் கருத்தில் கொண்டே எதிர்காலத்தில் ஒவ்வொரு இயல்களும் முன்வைக்கும் மொழிசார் புரிதல்களை வடிவமைக்க வேண்டும். தற்பொழுது அடிப்படை நிலையில் இருப்பதால் சிறு முயற்சி செய்து பார்க்கப்படுகின்றது. இம்முயற்சியில் இடம்பெறும் குறுஞ்செயலி உருவாக்கத் தன்மையை இனிக் காண்போம்.

படிநிலை-1:-

//Variable Declare

Button எழுத்ததிகாரம், சொல்லதிகாரம், பொருளதிகாரம்

எழுத்ததிகாரம்./ //*

**/*

முதல் படிநிலையாக ஒரு மாறியை உருவாக்கிக் கொள்ள வேண்டும். அதில் காட்சிப்படுத்த வேண்டிய 'எழுத்ததிகாரம், சொல்லதிகாரம், பொருளதிகாரம்' ஆகியவற்றைத் தரவேண்டும். அது பின்வருமாறு காட்சிப்படுத்தம் செய்யும்.

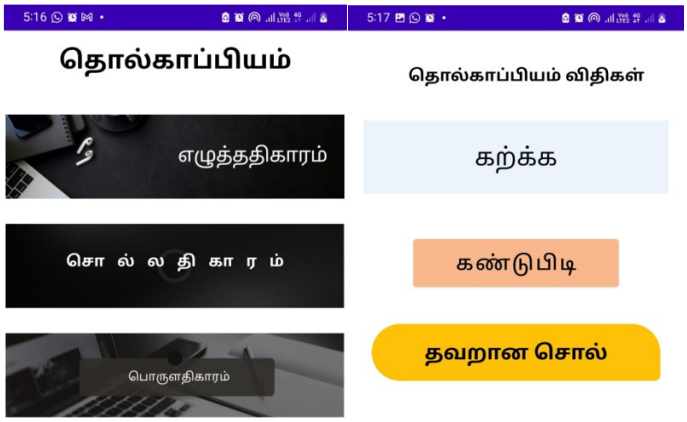
படிநிலை-2:-

```
Intent intent = new intent (Activity1.this, Activity2.class);

Startactivity(intent);
```

இந்தப் படிநிலை 2இன்படி எழுத்ததிகாரம் எனும் பகுதிக்குள் சென்று பின்வருமாறு காட்சிப்படுத்தம் நிகழும். பின்வரும் படங்கள் இக்குறுஞ்செயலி உருவாக்கத் தன்மைகளைக் காட்டும்.

இவ்வாறு அதன்படிநிலைகள் அமைந்து இந்தக் குறுஞ்செயலி உருவாக்கம் நிகழ்ந்து கொண்டிருக்கின்றது. இது ஒரு நீண்டகால ஆய்வு. அதன் முதல்படிநிலை ஆய்வுமுயற்சியை மட்டும் இவ்வாய்வில் கூறப்பெற்றுள்ளது.



நிறைவாக

தொல்காப்பியரின் தமிழ்மொழி சார்ந்த கருத்தியலாக்கங்களைப் பூதாகரமாக இல்லாமல் அனைவரும் பயன்படுத்தக்க வகையில் இக்காலத்தினர் பயன்பாட்டிற்கு ஏற்ப வடிவமைப்பெறல் வேண்டும் என்பதையும் அவ்வாறு உருவாக்குவது காலத்தின் தேவை என்பதையும் சாவா நிரல்மொழி கொண்டு இதுபோன்ற பல்வேறு தொழில்நுட்பச் சாதனங்களைக் கொண்டு வரமுடியும் என்பதையும் மேலே விளக்கப்பெற்ற கருத்தியல்கள் வெளிப்படுத்தி நிற்கின்றன.

துணைநின்றன

- புலியூர் கேசிகன், 2017, தொல்காப்பியம் (முழுவதும்), பாரிநிலையம், சென்னை.
- https://play.google.com/store/apps/details?id=com.cicttolkappiyameluttu
- https://play.google.com/store/apps/details?id=com.cict.tolkappiyam.col

Search Engines, Text Analytics, and Data Mining in 'Big Data'

K. Madhumita

In the era of the digital information explosion, the accumulation and management of vast amounts of data, often referred to as 'Big Data,' has become a central challenge and opportunity in various domains. The relentless growth of data sources from the web, social media, sensors, and traditional documents has prompted the development of advanced techniques in search engines, text analytics, and data mining to extract valuable insights, patterns, and knowledge from this sea of information.

While the field of 'Big Data' analytics has seen remarkable advances, much of the spotlight has focused on languages with well-established digital infrastructures and extensive language resources, such as English. However, the demand for the analysis and utilization of 'Big Data' in languages with comparatively limited resources has been on the rise. This paper embarks on an exploration of the intricate relationship between search engines, text analytics, and data mining in the realm of 'Big Data' with a particular emphasis on Tamil, a Dravidian language predominantly spoken in the Indian state of Tamil Nadu and Sri Lanka.

Tamil boasts a rich linguistic heritage, replete with a vast body of classical and contemporary literature, diverse cultural nuances, and a dynamic digital presence. This makes Tamil an intriguing subject for digital analysis, where innovative technologies could potentially unlock its immense potential. Yet, building effective search engines, text analytics tools, and data mining techniques for Tamil presents distinctive challenges owing to the complexity of its script, morphological richness, and limited digital footprint

This research paper addresses these challenges by closely examining the pivotal role of search engines, text analytics, and data mining in harnessing 'Big Data' for Tamil content. It aspires to offer deep insights into the adaptation and development of technology to facilitate the robust analysis of Tamil data, thereby contributing to the preservation and propagation of this ancient language in the digital age.

Introduction:

The potential locked within Tamil is immense, and innovative technologies have the capacity to unlock it. Yet, building effective search engines, text analytics tools, and data mining techniques tailored to Tamil presents a set of distinctive challenges. These challenges emanate from the language's complex script, morphological richness, and limited digital footprint, which altogether demand a specialized approach to 'Big Data' analysis.

This research paper aims to address these challenges by closely examining the pivotal role that search engines, text analytics, and data mining play in harnessing 'Big Data' for Tamil content. We aspire to offer deep insights into the adaptation and development of technology to facilitate the robust analysis of Tamil data. In doing so, our objective extends beyond technology; we seek to contribute to the preservation and propagation of this ancient language in the digital age, and by extension, to illuminate a path for similarly resource-limited languages to unlock their potential within the 'Big Data' landscape.

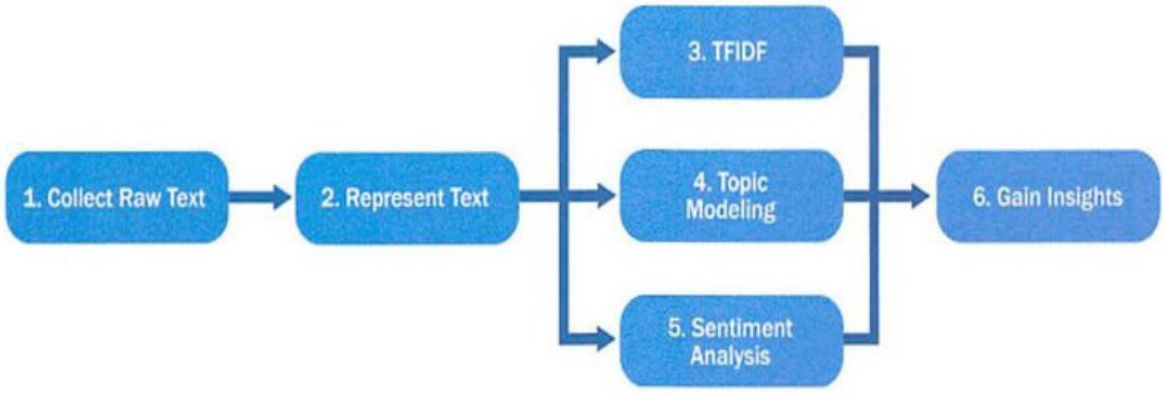
Theoretical Framework: Leveraging Technology for 'Big Data' Analysis in Resource-Limited Languages Information Retrieval and Search Engines:

• Information Retrieval Theory: Information retrieval (IR) theory forms the basis for understanding how search engines retrieve relevant data from vast corpora. The fundamental concepts of query processing, indexing, and relevance ranking underpin the functioning of search engines.
• Vector Space Models: The vector space model, which measures the similarity between queries and documents, is a key theoretical framework for search engine operations.

Corpus	Word Count	Domain	Website
Shakespeare	0.88 million	Written	http://shakespeare.mit.edu/
Brown Corpus	1 million	Written	http://icame.uib.no/brown/bcm.html
Penn Treebank	1 million	Newswire	http://www.cis.upenn.edu/~treebank/
Switchboard Phone Conversations	3 million	Spoken	http://catalog.ldc.upenn.edu/LDC97S62
British National Corpus	100 million	Written and spoken	http://www.natcorp.ox.ac.uk/
NA News Corpus	350 million	Newswire	http://catalog.ldc.upenn.edu/LDC95T21
European Parliament Proceedings Parallel Corpus	600 million	Legal	http://www.statmt.org/europarl/
Google N-Grams Corpus	1 trillion	Written	http://catalog.ldc.upenn.edu/LDC2006T13

Text Analytics and Natural Language Processing (NLP):

• **NLP and Linguistics**: The theoretical underpinnings of NLP, which includes grammar, syntax, semantics, and pragmatics, are essential to comprehend how text analytics tools process and analyse language.
• **Statistical NLP**: Statistical NLP techniques, such as n-grams and part-of-speech tagging, play a crucial role in text analytics. These theories help to extract meaning and structure from unstructured text data.



Data Mining and Machine Learning:

• **Supervised and Unsupervised Learning:** Understanding the theoretical foundations of supervised and unsupervised machine learning algorithms is essential for the development of data mining techniques. This includes regression, clustering, and classification.

- **Feature Selection and Engineering:** The theoretical concept of feature selection, extraction, and engineering guides the process of selecting relevant attributes for analysis and model building.

'Big Data' and Language Diversity:

- **Information Overload and Variety:** Theoretical perspectives on the challenges of handling 'Big Data,' which is characterized by volume, velocity, variety, and veracity, are relevant. In the context of resource-limited languages like Tamil, handling the variety of data sources becomes crucial.

- **Language Resource Scarcity:** The scarcity of linguistic resources for languages like Tamil raises theoretical questions about how to adapt existing technologies to accommodate these languages.

Methodology: Enabling 'Big Data' Analysis in Tamil

1. Research Design:

- **Exploratory Research: Given the evolving nature of 'Big Data' analysis in resource-limited languages, an exploratory approach will be adopted. This design allows for a flexible exploration of the complex relationships between technology and language.**

2. Data Collection:

- **Data Sources: A diverse range of data sources will be considered, including web content, social media, traditional documents, and user-generated content. This variety of sources is essential for comprehensive 'Big Data' analysis.**

- **Data Crawling and Collection Tools: Custom web crawlers and data collection tools will be developed to gather relevant data in Tamil from various online platforms.**

3. Linguistic Resources:

- **Tamil Language Resources: The methodology will leverage existing linguistic resources for Tamil, such as dictionaries, corpora, and language models, to enhance text analytics and data mining processes.**

- **Language Proficiency: Native Tamil speakers and linguistic experts will be involved in language verification and validation processes to ensure data accuracy.**

4. Preprocessing and Text Analysis:

• Text Cleaning: Raw data will undergo preprocessing, including tasks like data cleaning, normalization, and noise reduction.

• Text Analytics: Text analytics techniques, such as sentiment analysis, topic modeling, and entity recognition, will be applied to extract valuable information and patterns from the data.

5. Technology Adaptation:

• Search Engine Optimization: Search engines will be customized to handle Tamil language queries efficiently, considering the unique script and morphology.

• Natural Language Processing (NLP): NLP tools and algorithms will be adapted for Tamil, including stemming, tokenization, and syntactic parsing.

6. Data Mining and Machine Learning:

• Feature Engineering: Features specific to the Tamil language will be engineered to facilitate effective data mining.

• Classification and Clustering: Machine learning models will be developed for tasks like sentiment classification and content clustering in Tamil data.

7. Evaluation:

• Performance Metrics: The methodology will employ standard evaluation metrics for information retrieval, text analytics, and data mining tasks, with a focus on precision, recall, F1 score, and accuracy.

• User Feedback: User feedback and user experience testing will provide insights into the usability and effectiveness of the technology.

Discussion: Unleashing the Potential of 'Big Data' Analysis in Tamil

The preceding sections of this research paper have illuminated the intricate relationship between technology and language, specifically in the realm of 'Big Data' analysis for the Tamil language. Our discussion delves into the significance, implications, and potential avenues of this study.

Technology Adaptation and Language Complexity:

Our research has successfully demonstrated the adaptation of technology to accommodate the complexities of the Tamil language. The tailored search engines, text analytics tools, and data mining techniques have shown promise in effectively handling the unique features of Tamil, including its script, morphology, and linguistic richness. This adaptation has the potential to not only facilitate data analysis but also serve as a blueprint for adapting technology to other resource-limited languages.

Empowering Language Preservation:

One of the most profound implications of this research is its potential impact on the preservation and propagation of the Tamil language. Tamil, with its rich literary tradition and cultural depth, stands to gain substantially from 'Big Data' analysis. By enabling the efficient analysis of Tamil content, this technology empowers language preservation efforts and contributes to the continued vitality of this ancient language in the digital age. Furthermore, the digital representation of Tamil culture through this analysis has the potential to create a more profound connection between Tamil speakers, their language, and their heritage.

Data Mining and Cultural Insights:

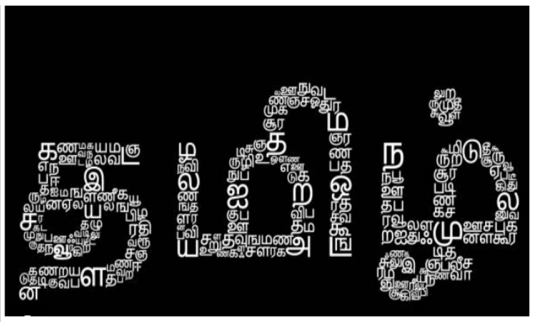
The data mining and content clustering techniques employed in this study offer more than just the capability to organize large datasets. They provide a unique lens through which to understand the cultural and sociolinguistic nuances of the Tamil-speaking community. The topical categorization of content enables researchers and cultural enthusiasts to explore and contextualize the diverse range of digital expressions in Tamil. This not only aids in academic pursuits but also fosters a deeper understanding of Tamil culture, both traditional and contemporary.

Understanding Text Analysis:

Text analysis, often referred to as text mining or natural language processing (NLP), encompasses a range of techniques and methodologies used to process, interpret, and extract meaningful information from unstructured textual data. In the context of 'Big Data,' it serves as the bridge that connects language to actionable insights. The primary objectives of text analysis in our research are to:

1. Content Organization: We employ text analysis to organize and categorize the vast and diverse content available in Tamil. Through techniques like topic modeling and content clustering, we aim to provide users with efficient access to relevant information.

2. Sentiment Analysis: Sentiment analysis is a crucial facet of our text analysis endeavors. By deciphering the sentiment expressed in Tamil text, we can gauge public opinion, emotional trends, and user reactions, thereby offering insights into the cultural and societal dynamics of theTamil-speakingcommunity.



Challenges in Text Analysis for Tamil:

While the importance of text analysis cannot be overstated, it is essential to recognize that adapting these techniques to the Tamil language presents a unique set of challenges. These challenges stem from the complexity of the Tamil script, its morphological richness, and the limited availability of linguistic resources. The following are key challenges in text analysis for Tamil:

1. Morphological Complexity: The rich agglutinative nature of Tamil presents a challenge in identifying word boundaries, stemming, and lemmatization. Traditional text analysis tools may require significant adaptation to handle the intricacies of Tamil morphology.

2. Script Variations: Tamil is written in multiple scripts, with the most common being the Tamil script (Brahmic) and the Tamil script (Indic). Text analysis tools must account for these variations to ensure accurate language processing.

3. Data Sparsity: Unlike English or other widely spoken languages, Tamil has limited linguistic resources, including dictionaries, corpora, and language models. This scarcity of resources necessitates innovative approaches to training and adapting text analysis tools.

Conclusion

Yet, as we conclude this research journey, we recognize that challenges remain. Data scarcity, script variations, and linguistic resource limitations are persistent hurdles in the 'Big Data' analysis of Tamil. These challenges necessitate continued collaborative efforts and resource development to enhance the accuracy and efficiency of technology adaptation.

In closing, our research endeavours are not merely a celebration of technology's triumph but also a testament to the enduring resilience of language. The Tamil language, with its depth of expression, will continue to thrive and evolve in the digital age, guided by our commitment to its preservation. As we adapt and develop technology to meet the unique needs of resource-limited languages, we create a blueprint for other languages to follow. Our journey is not an end but a beginning, and we look forward to future research endeavours that expand the horizons of 'Big Data' analysis and linguistic diversity.

As we embrace the transformative power of technology, we simultaneously embrace the enduring beauty of language. In the digital age, these two forces, intertwined, will propel us towards a future where every language, no matter how resource-limited, can unlock its digital potential and find its place in the global conversation.

தமிழ் மொழியில் செயற்கை நுண்ணறிவுப் பயன்பாடும் இன்றியமையாமையும்

முனைவர் த.புஷ்பராணி
உதவிப்பேராசிரியர்
தமிழ் இலக்கியத்துறை
நல்லமுத்துக்கவுண்டர்மகாலிங்கம் கல்லூரி, பொள்ளாச்சி.

முன்னுரை

இன்றைய காலகட்டத்தில் தொழில்நுட்பத்தைப் பொருத்தவரை, தொழில்நுட்பம் மட்டுமல்ல அனைத்துத் துறையினரும் கூறும் வார்த்தை "செயற்கை நுண்ணறிவு" என்பதாகும். ஏறக்குறைய இருபது, இருபத்தைந்து வருடங்களுக்கு முன்னர் "கணிப்பொறி" என்ற சொல் எவ்வாறு அனைத்துத் துறையினரையும் ஆர்வம் ஏற்படுத்தக் கூடிய வார்த்தையாக இருந்ததோ அதுபோல் இன்று செயற்கை நுண்ணறிவு அதாவது "AI"என்பது மாறி உள்ளது.

செயற்கை நுண்ணறிவு

கணிப்பொறி இன்று நம் வாழ்க்கையில் இரண்டறக் கலந்து விட்டது. கணிப்பொறி இன்றி நம்மால் இயங்க இயலாது என்னும் அளவிற்கு அனைத்துத் துறைகளிலும் நீக்கமற நிறைந்து இருப்பது கணிப்பொறி. அதன் படிநிலை வளர்ச்சியாக உருவெடுத்துள்ள ஒன்றுதான் செயற்கை நுண்ணறிவு என்பது ஆகும். கணிப்பொறி என்பது நாம் கொடுக்கும் உள்ளீடை பொறுத்து அதாவது எது சார்ந்த பணியை எவ்வாறு செய்ய வேண்டும் என நாம் அதற்கான மொழியின் மூலம் உள்ளீடு செய்கிறோமோ அவை அது போல் செயல்பட வேண்டுமென நாம் கூறுவதை வைத்து அவ்வேலையை விரைந்து துல்லியமாக முடித்துக் கொடுக்கக்கூடிய ஒரு கருவி. இது "Batch processing " என்பது மாதிரியான பணிகளுக்கு மிகச் சுலபமாக அமைந்தது. ஒரே வேலையை திரும்பத் திரும்ப செய்வதற்கு இது பேருதவி புரிந்தது. மனிதன் செல்ல முடியாத இடத்திற்குச் செல்வதற்கும் மனிதர்களால் செய்ய முடியாத பணியை செய்வதற்கும் மருத்துவம் முதற்கொண்டு தொழில்நுட்பவியல், கல்வியியல், நுண்ணணுவியல், எலக்ட்ரானிக்ஸ், கட்டிடவியல், மொழி இலக்கியம், விண்ணியல் என எல்லாத் துறைகளிலும் இக்கருவி பேருதவியாக இருந்து வருகிறது. அதன் பயன்களில் ஒரு பிரிவான ரோபோ போன்றவையும் மனிதர்களுக்கு இணையாக பல்வேறு பணிகளை செய்து வருகின்றன.

தமிழ் மொழி, இலக்கியம் சார்ந்து கணினி எவ்வாறு தொடக்க காலத்தில் நமக்குப் பல்வேறு விதங்களில் பயன்பட்டதெனில், பல்வேறு எழுத்து வடிவங்களையும் எழுத்து முறைகளையும் *(Fonts)* வசனம் உரையாடல் அமைப்புகளையும் *(Scripts)* மொழிபெயர்ப்பு சார்ந்தும் பெரிதும் உதவியாக இருந்தது வருகிறது. நிறைய இலக்கியங்களைச் சேமித்து வைக்கவும் தமிழ்த் துறை சார்ந்த அறிஞர்கள் வெவ்வேறு இடங்களில் இருந்தபோதும் அவர்களிடையே இருந்த மொழி, இலக்கியம் சார்ந்த கருத்துக்களைப் பதிவு செய்யவும், பரிமாறிக்கொள்ளவும் மற்றவர்களிடமிருந்து புதிய தகவல்களை அறிந்து கொள்ளவும் பல வழிகளில்*(Websites, Blogs)* உதவியாக இருந்தது.கலந்துரையாடல், கூட்டங்களில் பேசுதல் போன்றவற்றிற்கு கணினி சார்ந்த கருவிகள் பெரிதும் பயன்பட்டு வருகின்றன.

தற்காலத்தில் கணினி பயன்பாடு அனைத்து துறைகளிலும் பெரும்பங்காற்று வருவதற்கு சான்றாக ஒரு சிலவற்றை இங்கு நினைவு கூறலாம் அதாவது கட்டுமான துறையில் சில ஆண்டுகளுக்கு முன் வரை *"3D Designs "* என்பது மாதிரியான வடிவமைப்புகளில் மட்டும் இருந்த நிலை மாறி தற்பொழுது கட்டிடங்களே கூட *"3D"* அமைப்பு முறையில் உருவாக்குதல் என்பதான நிலை உருவாகி விட்டதை காண முடிகின்றது. கட்டிடங்கள் முழுவதையும் கணிப்பொறி அமைப்புகளே உருவாக்குவதையும் காணமுடிகின்றது. அனைத்து துறைகளிலும் கணினி சார்ந்த தானியங்கும் முறைக் கருவிகளும் பெருகி கொண்டு வருகின்றன. சான்றாக மருத்துவத் துறையிலும் ஒரு முறை தகவல்களை உட்செலுத்தி விட்டால் அறுவை சிகிச்சைகள் கூட கணினிகளே வெற்றிகரமாக செய்து முடிப்பதைக் காண முடிகின்றது.

தற்காலத்தில் தமிழ் மொழியில் செயற்கை நுண்ணறிவுப் பயன்பாடு

தமிழைப் பொருத்தவரை ஏதாவது ஒரு பிரிவு சார்ந்த நூல்களை உலகின் எந்த மூலையில் இருந்து கொண்டும் முழுமையாக அறிய இணையம் உதவுகின்றது. அதனுடன் செயற்கை நுண்ணறிவும் இணையும் பொழுது, ஒருவர் தொடர்ந்து வாசிக்கும் நூல்கள் எவை பற்றியன, அவரது வயது முதலான அடிப்படை விவரங்கள், அவரது தாய்மொழி போன்றவற்றுடன் அவர் நூல்களைப் படிக்கும் வேகத்தையும் உணர்ந்து அவருக்கு எத்தகைய நூல்களின் மீது விருப்பம், இனி எதிர்காலத்தில் எது தொடர்பான நூல்களில் கவனம் செலுத்தி விரும்பிப் படிப்பார், எத்தகைய எழுத்தாளர்களின் நூல்கள் அவரது ஆர்வத்திற்கு ஏற்ற வகையில் இருக்கும் என்பதை அறியும் வகையில் செயற்கை நுண்ணறிவு செயல்பட்டு அத்தகையவைகளை அவர் பார்க்கும், பயன்படுத்தும் *"Blog"* அல்லது *"Website"* ற்குக் கொண்டு வர ஆவண செய்ய வேண்டும். நாம் போரிடும் முறை இணையத்தில் தேடும் விவரங்களை கருத்தில் கொண்டு அவை போன்ற செய்திகளை அதிகமாக நம் பார்வையில் படும்படி இணையம் வழிவகை செய்கிறது அதேபோல் இலக்கியம் சார்ந்த நூல்களை நம் தேவைக்கேற்றபடி நமக்கு வழங்குகிறது செயற்கை நுண்ணறிவு முறைதான்.

இனி வரும் காலங்களில் இத்தகைய செயற்கை நுண்ணறிவின் மூலம் நூல்களை உருவாக்கவும் ஒரு வாசகர் எது தொடர்பான செய்திகளில், எந்தப் படைப்பாளரின் படைப்புகளில் அதிக ஆர்வம் செலுத்துகிறார் என்பதை உணர்ந்து, கணினி யே படைப்பாளியாகத் திகழும் வகையில் நலம் பயக்கும். தற்பொழுது நமக்கு ஏதேனும் செய்திகள் பிறரிடம் இருந்து வரும் போது அதற்கு நாம் அளிக்க வேண்டிய பதிலை இணையமே நமக்கு வழங்கி வருகிறது. எவருக்கு எத்தகைய வார்த்தைகளைப் பயன்படுத்த வேண்டும்? தொடர்கள் எந்த வகையில் அமைய வேண்டும்? என்பன போன்ற கருத்துக்களை செயற்கை நுண்ணறிவின் வாயிலாகவே கணினிகள் நமக்கு வழங்குகின்றன.

இலக்கியங்களை நூலகமாக இணையத்தில் நாம் சேமித்து வைத்திருப்பதைக் கொண்டு அதன் பலன் செயற்கை நுண்ணறிவாக நமக்கு கருத்துக்கள் வழங்கும் அளவிற்கு செயற்கை நுண்ணறிவு முன்னேறி உள்ளது என்பது குறிப்பிடத்தக்கது.

அடுத்ததாக, இளம் பருவத்தினரும் குழந்தைகளும் மிக விரும்பும் குரல் வழி அறிவிப்பை உணரும் திறன் கொண்ட கருவி *(Voice recognition, Speech recognition)*மிகப் பிரபலமானது. *" Alexa"* என்ற செல்லப் பெயர் இன்று குழந்தைகளின் மத்தியில் மிகவும் சிறப்புப்பெற்றுத் திகழ்கிறது. இது மனிதர்களை

இன்னும் சோம்பேறிகள் ஆக்கும் கருவியாகவும் சொற்களுக்கான எழுத்துக்களை அறியாதவர்களாகவும் ஆக்கும் இணையம் சார்ந்த கருவி என்றும் கூறலாம். *"Voice recognition tools, Speech recognition tools "* என்பவை குரல்றிதல் பேச்சு அறிதல் என்னும் கருவிகளாக உள்ளன இவற்றில் குரல் அறிதல் என்பதை *"Password "* ஆக கூட பலர் தங்களது கருவிகளில் பயன்படுத்துகின்றனர். அவர்களது குரல் மூலமாக அந்தக் கருவி *"Unlock"* ஆவதைக் காண முடியும்.

"Speech recognition"-ல் எல்லா வட்டார வழக்குகளும் இடம்பெற்றுள்ளன. அதாவது இலங்கைத் தமிழ், மலேசியத் தமிழ், இந்தியத் தமிழ் எனத் தற்போது இருக்கின்றன. இனிவரும் காலங்களில் இவை மதுரைத் தமிழ், கோயம்புத்தூர்த் தமிழ், பொள்ளாச்சித் தமிழ் என்பன போன்ற வட்டார வழக்குகளாக உருப்பெற்றாலும் வியப்பதற்கு இல்லை. ஏனெனில் இவையும் செயற்கை நுண்ணறிவின் மூலம் சாத்தியமே. இதனைக் கொண்டு அந்தந்த வட்டாரங்கள் சார்ந்த இலக்கியங்களை ஆராய சொற்றொடர்அமைப்பை உணர ஏதுவாகும். தமிழுக்கும் பிறமொழிகளுக்குமான தொடர்பு, தமிழுக்கும் பிற நாடுகளுக்கும் இடையேயான தொடர்பு பண்டைய காலங்களில் எவ்வாறு இருந்தது என்பதை உணரவும் இந்த செயற்கை நுண்ணறிவு பெரிதும் பயன்படுகிறது.

தமிழ் மொழியில் உள்ள சொற்களைப் பிற மொழிச் சொற்களுடன் ஒப்பிட்டுப் பார்த்து அச்சொல்லின் மூலமொழி, அதில் வழங்கப்பட்ட பொருள், அச்சொல் மற்ற மொழிகளில் எவ்வாறு திரிந்து வழங்கப்படுகிறது என்பன போன்ற தமிழ்மொழி சார்ந்த ஆராய்ச்சிகளை செயற்கை நுண்ணறிவைக் கொண்டு ஆராய முடியும். மனிதர்கள் இவ்வாறு அதிக எண்ணிக்கை அளவிலான மொழி அறிவைப் பெறுவதும் அவைகளை நினைவில் வைத்துக்கொள்வதும் தொய்வின்றி விரைந்து ஆராய்ந்து கண்டுபிடிப்பதும் இயலாத ஒன்று. எனவே இத்தகைய பணிகளுக்குச் செயற்கை நுண்ணறிவின் உதவி இன்றியமையாதது ஆகும்.தமிழ் மொழியில் உள்ள வார்த்தைகளுக்கான வரலாற்றையும் காலப்போக்கில் அவற்றின் மறுவிய வடிவங்களையும் அவை புழங்கப் பெறும் பிற மொழிகளையும் இவ்வகையில் ஆராயும் போது தமிழின் தொன்மைச் சிறப்பு இன்னும் மேலோங்கும்.

தமிழகத்தைப் பொறுத்தவரை கோயில்களில் சிற்பங்களும் கல்வெட்டுகளும் சிறப்பு பெற்றவை. இத்தகைய கல்வெட்டுகளில் உள்ள எழுத்துக்களின் உருவ அமைப்பைக் கொண்டு அவை எந்த நூற்றாண்டில் உருவாக்கப்பட்டவை? அதன் நோக்கம், அதனைக் கொண்டு அக்கால மக்களின் வாழ்வியல் போன்றவைகளை ஆராய தற்காலத்தில் முடிகிறது. தமிழகக் கல்வெட்டுகளில் உள்ள எழுத்துக்களைப் போன்றே தாய்லாந்து நாட்டில் உள்ள கல்வெட்டுகளிலும் காண முடிகின்றது. இவ்வாறு தமிழக மன்னர்கள் எந்தெந்த வேற்று தேசங்களுக்குச் சென்று அரசாட்சி செலுத்தினர்? அதன் மூலம் கலந்த, பரவிய பழக்க வழக்கங்கள், பண்பாடுகள் ஆகியவைகளை ஆராய செயற்கை நுண்ணறிவு உதவும்.

தற்பொழுது *"Call centres"* அனைத்துமே தானியங்கியாக மாற்றம் பெற்றுவிட்டது. இதன் மூலம் மனிதர்கள் கேட்கும் கேள்விகளுக்கு கனிவான பதிலளிக்க கூடியதாக கணினி செயல்பாடுகள் தற்பொழுது புழக்கத்தில் உள்ளன. இவற்றுள் செயற்கை நுண்ணறிவை உட்புகுத்தும் போது மனிதர்களே நின்று கொண்டுள்ளது போன்று செயல்பட்டு கோபமான, வருத்தமான பதிவுகள் வரும்போது அவற்றை முகத்திலோ, வார்த்தைகளிலோ பிரதிபலிக்காமல் கனிவுடன் பதில் தருவதோடு குரல் ஏற்றத்தாழ்வுடன் பாவனை இன்றி கூறுவதாகவும் செயற்கை நுண்ணறிவு உள்ளது குறிப்பிடத்தக்கது.

தமிழ் மொழியில் செயற்கை நுண்ணறிவின் எதிர்காலத் தேவை அனைத்துப் பிரிவுகளிலும் தகவல்கள் திரட்டுவதற்கு மிகவும் பயன்படுகிறது. இலக்கியம் சார்ந்த செய்திகள் நூல்வடியில் இருப்பதை ஒலி வடிவில் மாற்றி வழங்குகிறது. இதேபோன்று தேர்வுத் தாள்களில் உள்ள விடைகளை ஒலியாக மாற்றி வழங்கினால் உதவியாக இருக்கும். நேரம் சேமிக்கப்படும். விரைவான வாழ்க்கை சூழலில் இது மிகவும் பயனுள்ளதாக அமையும்.

தமிழில் *" Tongue Twister "* என்று சொல்லக்கூடிய நா பிறழ் வாக்கியங்கள் இலக்கியங்களில் அதிகமாக உள்ளன.இவைகளை தெளிவாகக் கற்றுக் கொடுக்கவும் உச்சரிப்பை திருத்தமாக்குவதற்கும் செயற்கை நுண்ணறிவை பயன்படுத்தலாம். சான்றாக திருக்குறளில் *"துறவு"* என்ற அதிகாரத்தில் அமைந்த இரண்டு குறட்பாக்களைச் சுட்ட முடியும். அதாவது நா ஒட்டாமல் உச்சரிக்கக் கூடிய குறட்பாவாக,

"யாதனின் யாதனின் நீங்கியான் நோதல் அதனின் அதனின் இல" (341) என்ற குறட்பாவை குறிக்கலாம்.

முழுவதும் நா ஒட்டிச் சொல்லக்கூடிய குறட்பா , "பற்றுக பற்றற்றான் பற்றினை அப்பற்றைப் பற்றுக பற்று விடற்கு" (350) என்பதாகும். இது போன்ற தொடர்களை ஆசிரியர்களின் துணை இன்றி, ஆனால் ஆசிரியர்களைப் போலவே கற்றுக் கொடுக்க செயற்கை நுண்ணறிவு அவசியமாகிறது.

"திருப்புகழ்" போன்ற இலக்கியங்களைப் பாடவும் இதன் இலக்கியச் செய்திகளை சந்த நடையில் கூறவும் இலக்கிய வளம் வாய்ந்த பல மனிதர்களாலேயே இயலாத ஒன்று. இத்தகைய பாடல்களில் உள்ள சந்தச் சிறப்பை எதிர்கால சந்ததியினர் அறிய ஏதுவாக இவற்றை செயற்கை நுண்ணறிவு கொண்டு கற்பிக்க வழிவகை செய்வது அவசியம் ஆகும்.

தமிழ் மொழியின் மொழியியல், இலக்கணம் ஆகியவை தமிழ் மொழியின் கருவூலங்கள் ஆகும். ஆனால் தற்காலத்தில் இவற்றை கற்கவும் கற்பிக்கவும் ஆர்வமும் முயற்சியும் குறைந்து வருவதைக் காண முடிகிறது. இந்நிலையை மாற்ற இவைகளைச் செயற்கை நுண்ணறிவுக்குள் உட்பகுத்தி அதன்வழி எளிமைப்படுத்திக் கற்கச் செய்யும் வழிமுறையும் தேவை. சான்றாக, ண, ந, ன: ல, ழ, ள: ர,ற: போன்ற எழுத்துக்களின் உச்சரிப்பு முறைகளில் பெரும்பாலும் பிழை ஏற்பட்டு வருகின்றன. இவை கொண்டு உருவாக்கப்படும் சொற்களும் வட்டார வழக்கிலும் மாற்றம் பெறுகின்றன. இவற்றின் அடிப்படையில் கற்பிக்கவும் இவற்றில் வரும் பிழைகளைச் சரி செய்யவுமான கற்றல் முறையை செயற்கை நுண்ணறிவின் வழி மேற்கொள்வது அவசியத் தேவை.

சான்று:
"Pipe"-குழாய் - கொளாய் -கீழ் – கொளா

கோழி – கோளி

முட்டை – மொட்டு

பழம் – பளொ போன்ற பல சொற்களைக் கூற முடியும்.

தமிழ் மொழியில் உள்ள "உரிச்சொற்கள்" எனப்படுபவை தமிழ் மொழியின் தொன்மை வளம் ஆகும். தொல்காப்பியத்தில் உரியியலாக தனி ஒரு இயல் அமைத்து தொல்காப்பியரால் தொகுத்துச் சொல்லப்பட்ட உரிச்சொற்கள் எனப்படுபவை அன்றாட வழக்கு சொற்களாக அல்லாமல் இலக்கியங்களில் மட்டுமே பயன்படுத்தக்கூடிய சொற்கள் ஆகும். இத்தகைய உரிச்சொற்கள் பல, சங்க இலக்கியங்களிலும் அதனைத் தொடர்ந்து வந்த பல வகை இலக்கியங்களிலும் காலம் காலமாக இடம் பெற்று வந்தன . ஆனால் தற்காலத்தில் இவை பெரும்பான்மை வழக்குழிந்து விட்டன. எவ்வகை இலக்கிய படைப்புகளிலும் பயன்படுத்தப்படுவதே இல்லை. இந்நிலை தொடருமே ஆனால் தமிழ் மொழியின் தொன்மை சிறப்பு வாய்ந்த உரிச்சொற்களை நாம் இழக்க நேரிடும் .எனவே இத்தகைய உரிச்சொற்களை மீட்டெடுக்க அவற்றைச் செயற்கை நுண்ணறிவுக்குள் கொண்டு வந்து உரிச்சொற்களுடனான வாக்கிய இலக்கியங்களை உருவாக்கினால் அது தமிழ் மொழிக்கு பெருஞ்சிறப்பு சேர்ப்பதாக அமையும்.

இன்றைய இளம் தலைமுறையினரின் தமிழ் எழுத்துக்களில் உள்ள பிழைகளைக் காணும் போது மிகுந்த வேதனை அளிக்கக் கூடியதாக உள்ளது. அதாவது பள்ளி, கல்லூரி மாணவர்களிடையே எழுத்துப் பிழைகள் ஏராளமாக மலிந்து கிடக்கின்றன. இதழ்கள் உள்ளிட்ட ஊடகங்களில் கூட பல்வேறு எழுத்துப் பிழைகள் காணப்படுகின்றன. இவ்வாறு பிழைகளுடன் மொழியை கையாளும்போது நம் கருத்தை எளிதாக மற்றவர்களுக்கு புரிய வைக்க இயலாத சூழல் ஏற்படும். மொழியை பிழையின்றி கையாளுவதே சிறப்பு. எனவே இத்தகைய பிழைகளைச் சரி செய்வதற்கான பயிற்சிகளை செயற்கை நுண்ணறிவின் துணைகொண்டு மேற்கொண்டோமே ஆனால் அது தமிழ் மொழியின் சிறப்பையும் வளத்தையும் காக்க நல்வாய்ப்பாக அமையும்.

தமிழ் நாட்டுப்புறவியல் என்பது தமிழர்களின் பண்பாட்டுக் கருவூலமாக திகழ்பவை எனலாம். இத்தகைய நாட்டுப்புற இலக்கியமானது பல்வேறு வகைகளாக உள்ளன. அவற்றில் நாட்டுப்புற பாடல்களே பல வகைகளில் பாடப்படுகின்றன. பிறப்பு முதல் இறப்பு வரையிலான அனைத்து

நிகழ்வுகளிலும் மனிதர்கள் பாடல்கள் மூலமாக தங்களது உணர்வுகளை வெளிப்படுத்துகின்றனர். இது பண்டைக்காலம் தொட்டு மரபாகப் பின்பற்றப்பட்டு வருகிறது. ஆனால் தற்காலத்தில் இத்தகைய பல பாடல் வடிவங்கள் வழக்கொழிந்து விட்டன. இவற்றை அந்தந்த முறை மாறாது நிலைத்திருக்கச் செய்ய வேண்டுமானால் அவைகளை செயற்கை நுண்ணறிவுக்குள் புகுத்தி அதன் வழி பரவச் செய்தல் அவசியம் ஆகும்.

தொடக்க காலம் தொட்டு இன்று வரை மனித வாழ்வானது சூழல் சார்ந்த வாழ்வாகவே உள்ளது இத்தகைய வாழ்வியல் முறையினை அந்தந்த கால இலக்கியங்கள் பதிவு செய்துள்ளன அவைகளின் வழி புழங்கு பொருட்கள் உணவு அவற்றை தயாரித்தல் அவைகளை பயன்படுத்துவதால் இயற்கைக்கு ஏற்படும் நன்மை போன்றவைகளை செயற்கை நுண்ணறிவுக்குள் புகுத்தி அத்துடன் கைவினைக் கலைகள்(மரத்தின் மூலம் ஏர் கலப்பை உருவாக்குதல், கூடை முடைதல், மரக்கட்டில்களுக்கான கயிறு பின்னுதல், கைவினைப் பொருட்களை உருவாக்குதல், இயற்கை உரங்களைத் தயாரித்தல், அவைகளைப் பதப்படுத்திப் பாதுகாத்தல், கூடை முனைதல், மாடுகளுக்கு மூக்கணாங்கயிறு போடுதல்) போன்றவற்றுடன் உணவே மருந்தான நம் சிறந்த வாழ்க்கை முறை போன்றவற்றையும் கட்டாயம் நிலைத்திருக்கவும் வளரச் செய்யவும் செயற்கை நுண்ணறிவின் உதவியால் மட்டுமே சாத்தியமாகும்.

அகராதியியல், தொல்லியல், சுவடியியல், சொற்பொருள் ஆராய்ச்சி, இலக்கணங்களை ஒளி - ஒலித் தொடர் படங்களுடன் *(Videos with audios)* விளக்கிச்சொல்லுதல் ஆகிய தமிழ்த் துறைகளுக்கு செயற்கை நுண்ணறிவின் பங்களிப்பு இன்றியமையாதது.

அவசர உலகில் வாழும் நம்மை, நம் மனதைப் புத்துணர்வூட்ட நம் பண்பாட்டுடன் இணைந்த கதைகள், பாடல்கள், நிகழ்த்து கலைகள் ஆகியவைகளையும் செயற்கை நுண்ணறிவின் வழி பரவச் செய்தால் இன்றைய இளம் தலைமுறைக்கு ஏற்படும் "மன அழுத்தம்" என்ற மாபெரும் இன்னலையும் அதனால் ஏற்படும் பல்வேறு தீய விளைவுகளையும் தடுக்க முடியும்.
இவ்வாறாக தமிழ் மொழிக்குள் நுண்ணறிவின் தேவையை எண்ணிப் பார்த்தோமானால் முடிவின்றி அவை தொடர்ந்து கொண்டே இருக்கும். ஆக மிக விரைவாகத் தமிழ் மொழிக்குள் செயற்கை நுண்ணறிவினை புகுத்துவது காலத்தின் கட்டாயமாகும்.

முடிவுரை

கல்தோன்றி மண் தோன்றா காலத்தே உருவான மூத்த குடியாம் நம் தமிழ்க் குடியின் பண்பாட்டையும் அதன் சிறப்புகளையும் உலகம் இருக்கும் அளவும் எடுத்து இயம்பவும் அதை உலகில் உள்ள பல்வேறு பிரிவினரும் கற்றுக் கொள்ளவும் மனிதர்கள் இருக்குமிடமெல்லாம் தமிழும் தமிழின் சிறப்புமிருக்க வேண்டும். அத்தகைய நிலைக்கு அவற்றைப் பரவ செய்வதற்கு மனிதர்கள் மட்டுமல்ல அவர்களுக்கு செயற்கை நுண்ணறிவின் உதவியும் அவசியமாகின்றது. எனவே தமிழைப் பொறுத்தவரை தமிழர்களை பொறுத்தவரை செயற்கை நுண்ணறிவின் தேவை என்பது உடலுக்கு சுவாசம் போன்று இன்றியமையாததாகும். எனவே அனைத்து துறைகளிலும் அதன் அனைத்து பிரிவுகளிலும் செயற்கை நுண்ணறிவை விரைவாகப் புகுத்துவதற்கு வழிகோலுவது அவசியமாகிறது.

செயற்கை நுண்ணறிவுத் தொழில்நுட்பமும் அதன் பயன்பாடும்

முனைவர் த.ராஜ்குமார்
இணைப்பேராசிரியர் & தலைவர்
தமிழ் இலக்கியத்துறை
நல்லமுத்துக் கவுண்டர் மகாலிங்கம் கல்லூரி
பொள்ளாச்சி
trajkumartamil@gmail.com

உலகின் அனைத்து உயிரினங்களைக் காட்டிலும் மனிதன் உயர்ந்திருப்பதற்கும் சிறந்திருப்பதற்கும் காரணம் சிந்தித்துச் செயல்படும் ஆற்றல் உடையவனாய் இருத்தலேயாகும். நெருப்பு மூட்டியது முதல் மின்சாரம் கண்டறிந்தது வரை தன் எண்ணம், சிந்தனைகளால் தனது தேவையை நிறைவேற்றிக் கொண்டான். விலங்குகளும் பறவைகளும் தனது ஆற்றலை வெளிப்படுத்துகின்றனவேயன்றி சிந்தனையை அல்ல. மனிதன் தன் தனித்த சிந்தனை ஆற்றலால் எந்தப் படைப்பையும் ஒன்று போல் அல்லாமல் தன் அறிவு காட்டும் வழி படைத்துக் காட்டுகிறான். ஐம்புலன்களின் மூலம் பெறப்படும் தகவல்களை ஆய்ந்து செயலாற்றி எதிர்காலத்தில் நடப்பனவற்றையும் சிந்திக்கும் திறனைப் பெற்றுள்ளான். தன் சிந்தனைத்திறனைக் கொண்டு மனிதனைப் போலவே செயல்படும் இயந்திர மனிதனைக் கண்டறிந்ததோடு அதையும் தாண்டி நுண்ணறிவோடு திறம்பட செயலாற்றும் செயலிகளையும் கணினியின் துணைகொண்டு கட்டுப்படுத்தும் அறிவையும் வெளிப்படுத்தி வருவதும் சிறப்புக்குரியதாகும்.

அறிவின் படிநிலைகள்

காலந்தோறும் மக்களின் அறிவியல், வானியல், சூழலியல் சார்ந்த அறிவுத்திறனை நம் இலக்கண, இலக்கியங்கள் பதிவு செய்துள்ளன. அவ்வகையில் அறிவின் நிலையை வகைப்படுத்திய தொல்காப்பியர்,

> *"ஒன்றறி வதுவே உற்றறி வதுவே*
> *இரண்டறி வதுவே அவற்றோடு நாவே*
> *மூன்றறி வதுவே அவற்றோடு மூக்கே*
> *நான்கறி வதுவே அவற்றோடு கண்ணே*
> *ஐந்தறி வதுவே அவற்றோடு செவியே*
> *ஆறறி வதுவே அவற்றோடு மனமே*
> *நேரிதின் உணர்ந்தோர் நெறிப்படுத் தினரே" (தொல்.மரபியல்,நூ.1526)*

என நெறிமுறைப்படுத்தியிருப்பதன் மூலம் மற்ற உலக உயிர்களைவிட சிந்தித்து செயல்படும் மனிதன் அறிவாற்றல் மிக்கவனாய் இருப்பதை உணரமுடிகிறது.

தொழில் நுட்ப மாற்றம்

எந்தவொரு தொழில்நுட்ப மாற்றமும் ஒரே நாளில் உருவாகிவிடுவதில்லை. தொழில்நுட்பத்தைப் பொறுத்தவரை கணினியின் வருகை, இணையப் பயன்பாடு, மின்னணுப் புரட்சி என இவற்றால் உலகம் இன்று ஒவ்வொரு வீட்டின் வரவேற்பறையிலும் சுழன்று கொண்டுள்ளது. இதற்கெல்லாம் மேலாக இன்றைய கணினி உலகம் செயற்கை நுண்ணறிவின் ஆளுகைக்குட்பட்டுள்ளது என்றால் அது மிகையில்லை.

கடந்த காலங்களில் மென்பொருள் ஆண்டுகொண்டிருந்த கணினி உலகத்தை இனி செயற்கை நுண்ணறிவு ஆளப்போகிறது. நம் அன்றாட வாழ்வில் நாம் பயன்படுத்தும் பொருட்களில் இந்நுண்ணறிவு செயல்பட்டுக்கொண்டுள்ளது. கண்காணிப்புக்கருவி அசைவை நோக்கி நகர்வதும், எந்த இடத்திற்கும் வழிகாட்டும் வரைபட வழிகாட்டியும், தேடுபொறிகளில் தேடக் கிடைக்கும் தரவுகள் அத்தனைக்கும் காரணமாக இருப்பது செயற்கை நுண்ணறிவே.

செயற்கை நுண்ணறிவு – விளக்கம்

செயற்கை நுண்ணறிவு என்பது ஒரு கணினி, கணினியால் கட்டுப்படுத்தப்படும் ரோபோ அல்லது ஒரு மென்பொருளை மனித மனதைப் போலவே புத்திசாலித்தனமாக சிந்திக்க வைக்கும் ஒரு முறையாகும்.

மனிதனின் பகுத்தறியும் திறனடிப்படையில், கற்றல், பகுத்தாய்தல், திட்டமிடல், உணர்தல், உள்ளுணர்தல், பார்த்தல், கேட்டல் ஆகிய பண்புகளைக் கொண்டு சூழ்நிலைக்கேற்ப முடிவுகள்

மேற்கொண்டு செயல்படுத்தக் கூடிய ஒரு பணியினைக் கணினியினைக் கொண்டு செய்து முடிக்க இயந்திரங்களை உருவாக்குவதாகும். எனவே, இத்துறை அனைவரின் கவனத்தையும் ஈர்த்து வருவதில் வியப்பேதுமில்லை.

செயற்கை நுண்ணறிவைக் கொண்ட இயந்திரம், மனிதர்களோடு சதுரங்கம் முதலான விளையாட்டுகளை விளையாடுகிறது; மருத்துவ சிகிச்சையளிக்கிறது. இதழியல் துறையில் இயல்பான மொழிநடையை உருவாக்கும் மென்பொருளானது தகவல்களைக் கொடுத்தால், கட்டுரையை குறைந்த நேரத்தில் உருவாக்கித் தந்துவிடுகிறது. வணிகத்தில் செயற்கை நுண்ணறிவைப் பயன்படுத்தி ஆளற்ற பல்பொருள் அங்காடிகள் உருவாக்கப்பட்டு வருகின்றன.

செயற்கை நுண்ணறிவுத் தாக்கம் பெற்ற துறைகள்

மின்னணு வணிகம்

மின்னணு வணிகம் அல்லது இணைய வர்த்தகத் தளங்கள் செயற்கை நுண்ணறிவின் தனிப்பயனாக்கக் காரணியைப் பயன்படுத்துவதற்கான மிகப்பெரிய தளங்களாக உள்ளன. இது சேகரிக்கும் தரவின் அடிப்படையில் பொருட்களை வாங்குவதற்கான பரிந்துரைகளின் பட்டியலை அளித்து உதவுகிறது.

சமூக ஊடகங்கள்

சமூக ஊடகங்கள் தற்போதைய தலைமுறையினருக்கு முதன்மையான பொழுதுபோக்குத் தளங்களாகும். இங்கு பகிரப்படும் இடுகைகள் மூலம் எண்ணற்ற தரவுகள் உருவாக்கப்படுகின்றன. எங்கெல்லாம் தரவுகள் நிறைந்திருக்கின்றனவோ அங்கெல்லாம் செயற்கை நுண்ணறிவு மற்றும் இயந்திர வழிக்கற்றல் ஆகியவை தொடர்பு கொண்டவையாக இருப்பதைக் காணமுடிகிறது.

கண்காணிப்பு

செயற்கை நுண்ணறிவானது பாதுகாப்புத் துறையில் பெரிய அளவில் வளர்ச்சியடைந்துள்ளது. தம் எல்லைக்குள் யாரும் அறியாத வண்ணம் ஊடுருவல் போன்ற பல அச்சுறுத்தல்களைக் கண்டறிவதில் பெரும்பங்காற்றி வருகின்றது.

மருத்துவம் - சுகாதாரம்

செயற்கை நுண்ணறிவு இத்துறையில் பெரும்பங்கு வகிக்கிறது. நோயாளிகளைக் கவனிப்பதில், மக்களுக்கு உதவுவதில் இது ஒரு முக்கியமான படியை எடுத்துள்ளது. சுகாதாரப் பயன்பாட்டு வசதிகளில் நோயாளிகளுக்கு முறையான மருந்து மற்றும் சிகிச்சையை உறுதி செய்யும் வகையில் இத்தொழில்நுட்பம் அமைந்துள்ளது.

சீனாவில் பல மருத்துவமனைகளில் மனித இயந்திரங்கள் மனிதர்கள் பணியாளர்களாக உள்ளன. மருத்துவமனைக்கு வரும் நோயாளிகளின் குரலையும் முகத்தையும் இனம் கண்டு அவர்களுக்கு உதவி வருகின்றன. மேலும் சீனமொழியின் வட்டார வழக்குகளையும் புரிந்து கொண்டு செயலாற்றுகிறது என்பது குறிப்பிடத்தக்கது.

வீடுகளில்

செயற்கை நுண்ணறிவு, வீடுகளில் பயன்படுத்தப்படும் மின்சாரப் பொருட்கள், பூட்டுகள், கதவுகள் போன்றவற்றிலும் வெளிப்புற வெப்பநிலையைக் கருத்திற்கொண்டு தானாகவே கட்டடத்தின் குளிர்ச்சி மற்றும் வெப்பத்தைக் கட்டுப்படுத்தும் திறனையும் வழங்குகின்றது.

வாகனங்களில்

செயற்கை நுண்ணறிவு அமைப்புகள் வாகனங்களின் நவீனக் கருவிகளின் வழியாகவும் புகைப்படக் கருவியின் வழியும் தரவுகளைச் சேகரித்து வாகனத்தை இயக்கும் கட்டுப்பாட்டு குறியீடுகளை வழங்குகின்றன. சில விலையுயர்ந்த வாகனங்கள் ஏற்கனவே செயற்கை நுண்ணறிவு உணர்திறன்களின் மூலம் இயங்கி வருகின்றன.

வேளாண்மைத்துறையில்

செயற்கை நுண்ணறிவு விவசாயிகளின் பணிச்சுமையைக் குறைக்கும் கருவியாக இல்லாமல் மாறாக அது அவர்களின் செயல்முறைகளை மேம்படுத்தும் பணியைச் செய்கின்றது. வெப்பநிலை, மழைப்பொழிவு, காற்றின் வேகம் மற்றும் சூரியக் கதிர்வீச்சு போன்ற விளைச்சலுக்கு உகந்த நுண்ணறிவுகளை அளித்து விவசாயிகளுக்கு உதவி வருகிறது. மேலும் காலநிலை மாறுபாடு, விளைச்சலைக் குறைக்கும் பூச்சிகள் மற்றும் களைகளின் தொற்று போன்ற சிக்கல்களைத் தீர்க்கும் சாத்தியக் கூறுகளையும் வழங்கி வருகின்றன.

மேலும் கல்வி, வானியல் ஆய்வுகள் உள்ளிட்ட பல்வேறு துறைகளிலும் செயற்கை நுண்ணறிவின் பங்களிப்பு இருந்து வருவது குறிப்பிடத்தக்கது.

செயற்கை நுண்ணறிவால் இயங்கும் செயலிகள்

தற்காலத்தில் செல்பேசிகளும் சமூக வலைதளங்களும் மனிதனின் உடலுறுப்புகளில் ஒன்று போல ஆகிவிட்டன. பயன் அடிப்படையில், 'செல்' இல்லாமல் எங்கும் செல்வதில்லை என்ற நிலை உருவாகிவிட்டது. தான் செய்யும் அத்தனை வேலைகளிலும் கணினி சார்ந்த பயன்பாடு இருப்பது அவ்வேலையை எளிதாக்குகிறது. எனவே அனைவரும் அன்றாடவாழ்வில் செயலிகளைக் கொண்டே செயலாற்றிக் கொண்டுள்ளனர் எனலாம். இவ்வாறு மக்கள் பயன்படுத்தும் செயலிகள் அனைத்தும் செயற்கை நுண்ணறிவின் துணையோடு திறம்பட செயலாற்றுகிறது.

- *chatgpt - https://chat.openai.com/*
- *bing - https://www.bing.com/*
- *facetune - https://www.facetuneapp.com/*
- *lensa - https://prisma-ai.com/lensa*
- *alexa - https://www.ultimatevoiceassistant.com/*
- *siri - https://www.apple.com/in/siri/*
- *Cleo - https://web.meetcleo.com/*
- *Google assistant - https://assistant.google.com/* இன்னும் பிற. . .

ஒன்றோடு ஒன்று போட்டி போட்டுக்கொண்டு வளரும் பல கோடி மதிப்பிலான இணையவழி பயன்பாட்டுச் சந்தையானது செயற்கை நுண்ணறிவை ஏற்றுக்கொள்வதற்கான மையமாக மாறியுள்ளது. ஒரு பயனர் புகைப்படத்தைத் திருத்த விரும்பினாலும், புதிய மொழியைக் கற்க விரும்பினாலும் அல்லது தொலைபேசி அழைப்பை எழுத விரும்பினாலும் இப்படியான பயன்பாட்டில் செயற்கை நுண்ணறிவுச் செயலிகள் தவறாமல் இடம் பெற்றிருக்கிறது.

சாட்ஜிபிடி

சாட்ஜிபிடி ஆனது பயனர்களுடன் உரையாடவும், கொடுக்கப்படும் வினாக்களுக்கு விடையளிக்க, புதிய உரையை உருவாக்க ஒரு கருவியாகப் பயன்படுத்தப்படுகிறது.

பிங்

மைக்ரோசாஃப்ட் பிங் நீண்ட காலமாக இணையத்தை ஆராய்வதற்கான தேடுபொறியாக அறியப்படுகிறது, ஆனால் கேள்விகளுக்கு பதிலளிக்கவும், ஆக்கப்பூர்வமான உரை மற்றும் படங்களை உருவாக்கவும், பல மொழிகளில் எழுதுவதற்கும் மொழிபெயர்ப்பதற்கும் சரிபார்ப்பதற்கும் பயன்படுத்தப்படுகிறது.

ஃபேஸ்ஆப்

இந்த செயலி பயனர்களின் புகைப்படத்தை மேம்படுத்த உதவுகிறது. முகத்தின் ஒப்பனை, சிகை அலங்காரம் போன்ற அழகியலை தரப்படுத்தப் பயன்படுத்தப்படுகிறது.

லென்சா

இதுவும் பயனர்களுக்கு புகைப்படங்களில் கலைத் திருத்தங்கள் மற்றும் மறு செய்கைகளை உருவாக்கும் திறன்களை வழங்குகின்றது.

அலெக்சா

2014 இல் அறிமுகப்படுத்தப்பட்டதிலிருந்து, அமேசானின் அலெக்சா ஒரு வீட்டுப் பணியாளராக மாறியுள்ளது. அதன் அதிநவீன இயற்கை மொழி செயலாக்கத் திறன்கள் அலெக்சாவை பேசும் மொழியைப் புரிந்துகொள்வது மட்டுமல்லாமல், பயனர்களுடன் எளிமையாக உரையாடவும் அனுமதிக்கின்றன.

***Google* உதவியாளர்**

இது மிகவும் மேம்பட்ட மெய்நிகர் உதவியாளர்களில் ஒன்றாகக் கருதப்படுகிறது. கண்ணுக்குப் புலனாகாத மனிதனைப்போல் நம்முடன் உரையாடி உதவிசெய்கிறது. நாம் கூறுகின்ற கட்டளையை ஏற்று அதன்படி செய்யும். நாம் சொல்வதைக் கேட்பதால், இங்கிவனை யான் பெறவே என்னதவம் செய்துவிட்டேள் என பாரதி கூறுவதுபோல நாமும் மெச்சிக்கொள்ளலாம்.

மேற்கண்ட செயலிகள் அனைத்தும் மக்களுக்கு நாள்தோறும் பயனளிக்கும் வகையில் அமைந்துள்ளன.

பிழை திருத்திகள் - ஆங்கிலம்

பிழையில்லாமல் எழுதுவதுதான் சொல்ல வரும் கருத்தை தெளிவாக உணர்த்துவதற்குரிய வழியாகும். பிழையாக எழுதினாலும் அதைத்திருத்தம் செய்யும் மென்பொருட்கள் அப்பணியைச் சிறப்பாகச் செய்யவேண்டும். பெரும்பான்மையர், பிழையின்றி எழுதும் ஆர்வம் இருப்பினும் அறியாமையாலும் ஐயத்தாலும் தவறாக எழுதி விடுகின்றனர். இவர்களுக்கு வழிகாட்ட வேண்டியது கணிப்பொறி வல்லுநர்களின் கடமையாகும். ஆங்கிலமொழி பிழைத்திருத்திகள் கீழ்வருமாறு.

Grammarly - https://www.grammarly.com/

ஆங்கில மொழியில் தயாரிக்கும் ஆவணங்கள், மின் அஞ்சல்கள், முகநூல், இன்ஸ்டாகிராம், ட்விட்டர் பதிவுகள் என அனைத்தையும் திருத்தி அமைக்கும். எம்.எஸ்.வேர்டு சுட்டிக் காட்டும் பிழைகளைக் காட்டிலும் பத்து மடங்கு அளவில் பிழைகளை இது சுட்டிக் காட்டுகிறது. முன் ஒட்டுச் சொற்கள் *(preposition)*, சரியான வினைச்சொல் பயன்பாடு, பெயர்ச்சொல் பயன்படுத்தல் தவறாகப் பயன்படுத்தப்படும் சொற்கள் என அனைத்து வகைகளிலும் இது நமக்கு உதவியாக இருக்கிறது.

Ginger Grammar Checker - http://www.gingersoftware.com/

இணையத்தில் கிடைக்கும், ஆங்கில மொழிப் பயன்பாட்டினை மெருகூட்டித் தரும் ஒரு கருவி 'ஜிஞ்சர்' *(Ginger)*. மேலே சொல்லப்பட்ட 'கிராமர்லி' போலவே, இலவசமாகவும், கட்டணம் செலுத்தியும் இதனைப் பெறலாம். இலக்கண மற்றும் சொல் பிழைகளை இது திருத்துகிறது.

Paper Rater - https://www.paperrater.com/free_paper_grader

இணையத்தில் கிடைக்கும் ஒரு வித்தியாசமான ஆங்கில மொழி திருத்தியாகும். இந்த செயலியும் எழுத்து, இலக்கண, சொல் மற்றும் வாக்கியப் பிழைகளைத் திருத்துகிறது. ஆங்கில மொழியினைப் பயன்படுத்துவதில், எழுதுபவர் அல்லது படிப்பவரின் நிலைக்கேற்ப, உரையைத் திருத்தம் செய்கிறது. கல்லூரி மாணவர், பட்டதாரி, முனைவர் பட்ட ஆய்வாளர் என மூன்று நிலைகளில் எப்படி ஆங்கிலம் பயன்படுத்தப்பட வேண்டுமோ அந்த நிலைக்கேற்ப திருத்தங்களை அளிக்கிறது.

After the Deadline - http://www.polishrtywriting.com

இலக்கணப் பிழைகளைத் திருத்தம் செய்து எவ்வாறு பயன்படுத்த வேண்டும் என அறிவுரை வழங்குகிறது.

WebSpellChecker - https://www.spellchecker.net/

இந்த செயலியின் சிறப்பு, ஒரு சொல்லுக்கு அதே பொருளைத் தரும் இன்னொரு சொல்லைத் தரும் எழுத்துப் பிழைகளைத் திருத்துவதுடன், இலக்கண பிழைகளை ஆய்வு செய்து விளக்கங்களைத் தருகிறது.

Online Correction - http://www.onlinecorrection.com/

இந்த *Online Correction* கருவி உங்களின் ஆவணத்தில் எத்தனை பிழைகள் உள்ளன என்று காட்டி, அவற்றை எப்படித் திருத்த வேண்டும் என்பதைச் சுட்டிக் காட்டுகிறது சொற்கள், வரிகளுக்கிடையே இடைவெளி சரியாக உள்ளதா என்பதையும் எடுத்துச் சொல்கிறது எழுத்துப் பிழைகள் சிகப்பு கோடு இடப்பட்டுக் காட்டப்படுகிறது. மற்ற பிழைகள், பச்சை வண்ணத்தில் கோட்டுடன் சுட்டிக் காட்டப்படுகிறது.

மேற்கண்ட பிழைத்திருத்திகள் அனைத்தும், ஆங்கில மொழியில் அமைக்கும் ஆவணங்களைத் திருத்தி, சிறப்பாக, உயர்ந்ததாக அமைக்க நமக்கு உதவுகின்றன.

பிழை திருத்திகள் - தமிழில்

ஆங்கிலத்தில் தட்டச்சு செய்தல், சரிபார்த்தல், தொகுத்தல் போன்ற பணிகள் கணினியின் மூலம் எளிதாக நிறைவேறிவிடுகிறது. ஆங்கிலத்தில் உள்ளது போலப் பிழை திருத்தும் *(spell checker)* வசதி தமிழில் இல்லாதது பெருங்குறையாயிருந்தது. தற்போது தமிழில் சில பிழைதிருத்தி மென்பொருள்கள் கிடைத்துவருகின்றன.

நாவி தமிழ்ச் சந்திப்பிழை திருத்தி

இந்த மென்பொருள் செயலியில் வாக்கியத்தைப் போட்டு *"ஆய்வு செய்"* பொத்தானைத் தட்டினால். வலிமிகும் இடங்களை ஆராய்ந்து பரிந்துரைக்கும். அதுபோகக் கணிக்கமுடியாத வார்த்தைகளுக்கு ஏற்ற இலக்கண விதிகளைச் சுட்டிக்காட்டும். வலி மிகாத இடங்களில் தவறாக வலி மிகுந்தாலும் கண்டுபிடித்துக் காட்டும்.

வாணி தமிழ் எழுத்துப்பிழை திருத்தி

இணையவழியில் அனைவரும் பயன்படுத்தும் வண்ணம் வடிவமைக்கப்பட்டு வாணி என்ற தமிழ் எழுத்துப்பிழை திருத்தி உருவாகியுள்ளது. நாவியில் பயன்படுத்தியது போல வாக்கியங்களை வாணியில் கொடுத்து *"திருத்துக"* பொத்தானை அழுத்தவும். பின்னர் பிழை திருத்தியபிறகு *"சம்மதம்"* பொத்தானை அழுத்தி, திருத்தங்களைப் பெற்றுக் கொள்ளலாம்.

மென்தமிழ்

இச்செயலி தமிழைப் பிழையற எழுத உதவும் மென்பொருளாக விளங்குகிறது. இதனை, தமிழ்த் தட்டச்சு விசைப்பலகைகள் ***(Tamil Keyboards)***, ஒருங்குறி எழுத்துருக்கள் ***(Unicode Fonts)***, குறியேற்ற மாற்றி ***(Encoding Converter)*** சொற்பிழை திருத்தி ***(Spell Checker)***, சந்திப்பிழை திருத்தி ***(Sandhi Checker)***, தமிழ்ச்சொல் சுட்டி ***(Tamil Word Suggester)***, அகராதிகள் ***(Dictionaries)***, அகரவரிசைப்படுத்தம் ***(Sorting)***, சொல்லடைவு ***(Indexing)***, துணைநூற்பட்டியல் கருவி ***(Bibliography)***, எண்<->எழுத்து மாற்றி ***(Number to Word Converter)*** போன்ற பல வசதிகளுடன் இச்செயலி அமைக்கப்பெற்றுள்ளது.

இவ்வகையான பல பிழை திருத்திகள் தமிழ்க் கணினி உலகிற்கு வந்து கொண்டிருக்கின்றன. எனினும் பயனீட்டாளர் தேவையை இவை ஓரளவு நிறைவு செய்வதாக அமைகின்றன. இன்னும் இத்துறை வளரவேண்டும். இலக்கணப் பிழைத் திருத்தி என்ற நிலையில் பிழை திருத்திகள் உருவாக வேண்டும். அவற்றை எதிர்நோக்கிக் கணினி பயன்படுத்துனர் காத்திருக்கின்றனர்.

செயற்கை நுண்ணறிவின் நன்மைகள் மற்றும் தீமைகள்

அறிவியல் வளர்ச்சி மனித வாழ்வை மேம்படுத்துவதாக இருந்தாலும் புதிய கண்டுபிடிப்புகளால் நன்மை இருப்பதைப்போலவே தீமையும் இருக்கத்தான் செய்கிறது. அவ்வகையில் செயற்கை நுண்ணறிவு பல நன்மைகளைத் தரும் அதே வேளையில் மனிதனை சோம்பேறி ஆக்கி விடுமோ என்ற கவலையும் எழுவதாக ஆய்வாளர்கள் பலரும் முன்வைக்கும் கருத்தாக உள்ளது,

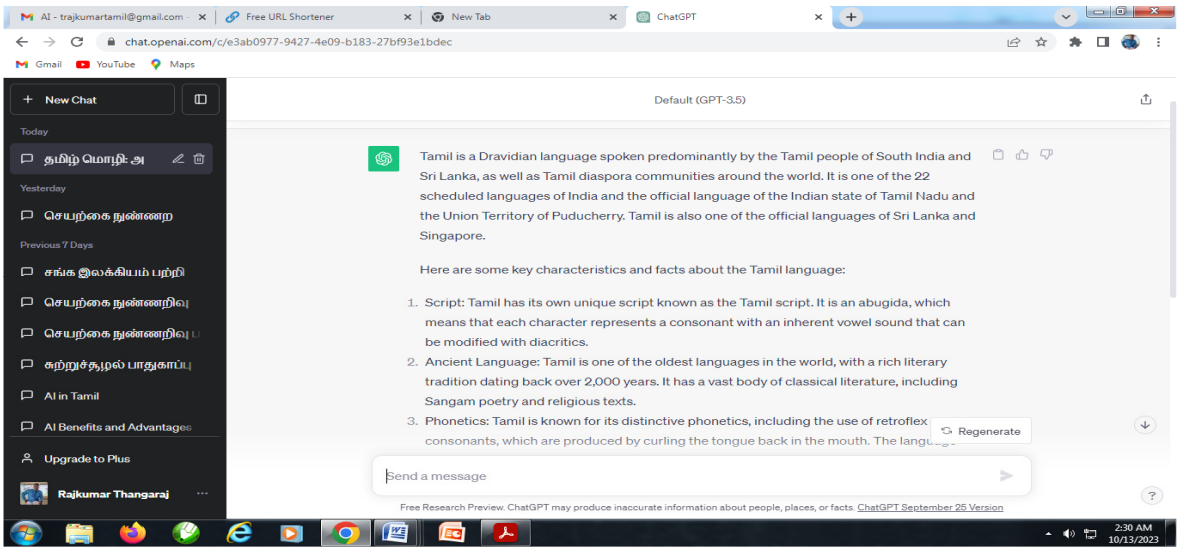
- செயற்கை நுண்ணறிவின் நன்மைகள்
  - பிழையற்ற செயலாக்கம்
  - மீண்டும் மீண்டும் வேலைகளில் உதவுகிறது
  - 24/7 கிடைக்கும்
  - சரியான முடிவெடுத்தல்
  - டிஜிட்டல் உதவி
  - வேகமாக முடிவெடுத்தல்
  - ஆபத்தான சூழ்நிலைகளில் செயற்கை நுண்ணறிவை செயல்படுத்துதல்
  - புதிய கண்டுபிடிப்புகள்
- செயற்கை நுண்ணறிவின் தீமைகள்
  - உருவாக்கத்திற்கான அதிக செலவுகள்
  - அதிகரித்த வேலையின்மை
  - படைப்பாற்றல் இல்லாதது
  - முன்னேற்றம் இல்லாதது
  - மனித பிரதிபலிப்பு இல்லை

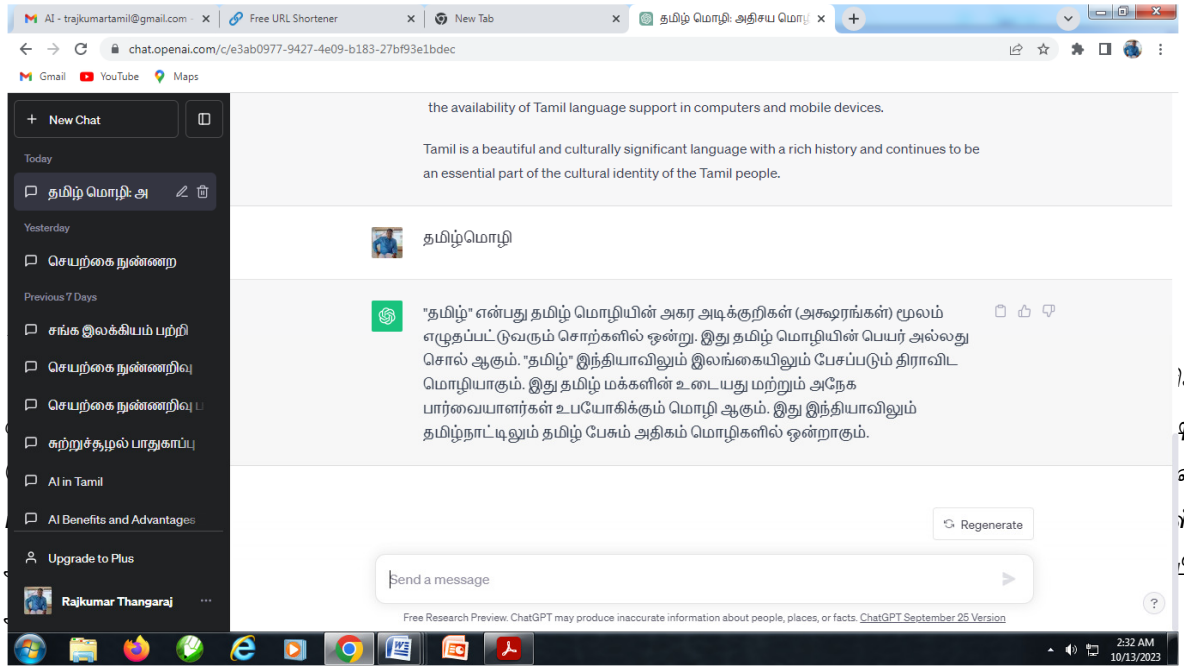
இருந்த போதும் நன்மையே மேலோங்கி இருக்கும் செயற்கை நுண்ணறிவின் பயனை மறுப்பதற்கில்லை.

தமிழ்மொழிப் பயன்பாட்டில் செயற்கை நுண்ணறிவு

செயற்கை நுண்ணறிவு தமிழ் வழியில் செயல்படுவதைக் காட்டிலும் ஆங்கில வழியில் சிறப்பாக செயல்பட்டு வருகின்றது. தமிழ் மென்பொருள்களின் பயன்பாடும் எண்ணிக்கையும் பெருகியிருந்தாலும் செயற்கை நுண்ணறிவு கொண்டு செயல்படும் செயலிகளில் தமிழ் மொழியில் வினவப்படும் வினாக்களுக்கு போதுமான தரவுகள் இன்மை, உச்சரிப்புகள், ஒலிக்குறிப்புகள் தவறாக வெளிப்படுவதும் காணமுடிகின்றது. எனவே கணினித்துறை சார்ந்த வல்லுநர்கள் இதற்காக முயலவேண்டும்.

கீழ்க்காணும் படங்கள் இரண்டும் *chatgpt* யில் தமிழ்மொழி என்னும் பொருளில் வினவப்பட்டவை. ஆங்கிலத்தில் தமிழ்மொழி குறித்த கருத்துக்களுக்கும் தமிழில் தமிழ்மொழி குறித்த கருத்துக்களுக்கும் எவ்விதத் தொடர்பும் இல்லாமல் இருப்பதைக் காணமுடிகின்றது. எனவே செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தை தமிழ்மொழியில் சிறப்பாகச் செயல்படுத்துவதற்கான முன்னெடுப்புகளைச் செய்வது அவசியமாகும்.





துணை நின்றவை :

இணையதளங்கள்

- http://manidal.blogspot.com/2017/08/blog-post.html
- https://builtin.com/artificial-intelligence/ai-apps

- *https://rb.gy/og2xe*
- *https://shorturl.at/oprTV*
- *https://www.skillrary.com/blogs/read/use-of-artificial-intelligence-in-agriculture*
- *X th Tamil Book*

Linguistic Translator

Vidhya Kanagaraj
KG College of Arts and Science

Machine translation, the automated conversion of text from one language to another, has witnessed remarkable advancements in recent years, with a growing focus on bridging the linguistic diversity of India. This work presents an overview of the challenges and advancements in machine translation systems designed to facilitate seamless communication between Tamil and English, as well as between various Indian languages.

linguistic diversity in India is characterized by a multitude of languages and dialects spoken across the nation. Tamil and English, being prominent among them, present unique challenges due to their linguistic differences in terms of grammar, vocabulary, and syntax. This abstract delves into the current state-of-the-art approaches employed in machine translation for Tamil-English language pairs, highlighting the application of neural machine translation (NMT) models, deep learning techniques, and the utilization of large parallel corpora to improve translation quality.

Furthermore, the work explores the broader scope of machine translation for inter-Indian language communication. It discusses efforts to develop translation systems between various Indian languages, considering languages like Hindi, Bengali, Telugu, and more. These initiatives not only aim to enhance linguistic accessibility but also promote cultural exchange and economic cooperation within India's diverse linguistic landscape.

The work also touches upon the challenges of translating languages with varying scripts, morphologies, and idiomatic expressions. Additionally, it discusses the importance of domain-specific translation models to cater to diverse sectors like healthcare, education, and e-commerce.

This abstract provides an insight into the evolving landscape of machine translation for Indian languages, with a particular focus on Tamil, English, and inter-Indian language translation. It underscores the importance of continued research and development in this field to foster cross-cultural communication, economic growth, and knowledge dissemination within the Indian subcontinent.

Promoting linguistic accessibility and cultural exchange. Challenges related to different scripts, morphologies, and domains are discussed, broader scope of inter-Indian language translation.

Extractive Summarization of Text Document

M.Shree Gowri[1],V.Srividhya[2]
Final Year MCA[1],Assistant Professor[2]
[1,2]Department Of Computer Science,Avinashilingam Institute for Home Science and Higher Education for Women ,Coimbatore India
shreegowri121@gmail.com[1],vidhyavasu@gmail.com[2]

Abstract

In an era characterized by an exponential growth of information on the World Wide Web, the need for effective text summarization techniques has never been more pressing. Our research addresses this imperative by focusing on the task of extractive summarization specifically tailored for text documents. Text summarization, which falls under AI, has been an important research area that identifies the relevant sentences from a piece of text. This paper presents an **"Extractive Summarization of Text Documents**" approach for generating a short and precise extractive summary for the given text documents. It is crucial due to the growing volume of text data, and it can be categorized into extractive and abstractive methods. In this paper, an extractive approach is employed, involving the selection and concatenation of important sentences or paragraphs from the original text. The objective is to provide concise summaries, which is a challenging task in natural language processing. It comprises four phases: The first phase of the work is loading the input text document. Text preprocessing is done in second phase. The third phase is applying extractive summarization algorithms such as frequency summarizers, Textrank, and Lexrank. The fourth phase of the work is performance evaluation using rough scores .

Keywords : extractive Summarization, rouge scores,Textrank.

Tamil - Ai Powered Legal Documentation Assistant

Dr. Karthikeyan Viswanathan[1],
Mr. Nithin.Y.J [2], Mr. Krishna.P.G[3], Mr. Prasanna.V[4], Mr. Faiz Alam.F[5], Mr. Nitheshwaran K[6]
Associate Professor[1], Pre-Final year[2,3,4,5,6]
[1,2,6] Department Of Mechanical Engineering, Sri Krishna College of Technology, Coimbatore, India
[3,4,5] Department Of Computer Science Engineering, Sri Krishna College of Technology, Coimbatore
Karthickeyan.v@skct.edu.in [1], 727821tume103@skct.edu.in [2], cseskct153prasanna.v@gmail.com [4]

ABSTRACT

Artificial intelligence is expected to revolutionize the legal profession by becoming an indispensable tool in the creation and management of legal documents. Legal documentation can be complicated and a time consuming process, especially for the individuals, who may not have access to legal resources and find it difficult to understand and follow-up. The language and jargon used in legal documentation are high in vocabulary and can lead to errors and misunderstandings. Tamilians with speech and hearing impairments face higher possibilities of difficulties in understanding and dealing with legal documentations. The objective is to develop an AI – powered solution that can simplify and translate the documentation into Tamil, by automatically drafting legal documents in plain language and using easy to understand terms. An App or a website that can analyze, review and give translation to complicated terms in the legal documents to the people for better understanding, powered with AI-algorithm.

An User-Friendly Interface for inputting relevant information such as parties involved, terms and conditions of the agreement and other necessary information. AI – powered documentation generation that automatically drafts the legal documents as per the input data given and also translates the legal terms and documents in Tamil. It has the ability to resolve the queries and doubts of the users in an efficient manner. Integration with existing legal resources and database to ensure accuracy and completeness of the legal documents. Option for the user to seek the end results in descriptive form, voice assistant and ASL sign language.

STUDIO
KCT
Explore..Capture..Inspire..

KUMARAGURU

ENGINEERING | LIBERAL ARTS | AGRICULTURE | MANAGEMENT
BUSINESS INCUBATION | TAMIL RESEARCH

KUMARAGURU COLLEGE OF TECHNOLOGY
Character is life

KCT BUSINESS SCHOOL

KCLAS KUMARAGURU COLLEGE OF LIBERAL ARTS AND SCIENCE

KUMARAGURU INSTITUTE OF AGRICULTURE

KUMARAGURU SCHOOL OF BUSINESS

KRiA KUMARAGURU RESEARCH & INNOVATION ALLIANCE

KUMARAGURU SCHOOL OF INNOVATION

KUMARAGURU KC.IRI CENTRE FOR INDUSTRIAL RESEARCH & INNOVATION

FORGE

SEA SAKTHI EXCELLENCE ACADEMY
Career Integrated Education

N.MAHALINGAM TAMIL RESEARCH CENTRE

KCT TECH PARK

KARE Kumaraguru Action For Relief And Empowerment

Dr.N.MAHALINGAM CHESS ACADEMY

CHARACTER IS LIFE

25 UG, 20 PG, 12 R&D Centres, 75+ MoUs, 71 Patented Projects, 60 Student Clubs & Forums, 182+ Laboratories, 900+ Faculty & Staff, 8000+ Students, 29,000+ Alumni, 1 Lakh+ Books in the Library, 100 Startups Incubated, 60+ Products/ technologies commercialised at Forge, 80,000 Scholastic Tamil Books in NMTRC, 8 MNCs in KCT Tech Park, 292 Acres of three Sustainable Campuses with 6000+ trees of 100+species, Resource Recovery Park.