ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

नैसर्गिक भाषा प्रक्रियेच्या माध्यमातून भारतीय भाषांचे संवर्धन

डॉ. अतुल मधुकरराव देशमुख
साहाय्यक प्राध्यापक व मराठी विभागप्रमुख, सौ. रजनीताई नानासाहेब देशमुख कला, वाणिज्य व विज्ञान महाविद्यालय, भडगाव, जि. जळगाव.
ईमेल: atuldeshmukh156@gmail.com

सारांश:
भारतीय भाषिक परंपरा ही जगातील सर्वात समृद्ध, बहुभाषिक आणि बहुसांस्कृतिक वारसा प्रणालींपैकी एक मानली जाते. भारतीय राज्यघटनेत २२ अनुसूचित भाषा असून शेकडो प्रादेशिक बोलीभाषा आणि आदिवासी भाषा अस्तित्वात आहेत. तथापि, जागतिकीकरण, शहरीकरण, इंग्रजीचे वाढते वर्चस्व, तंत्रज्ञानातील असमान प्रवेश, शिक्षणातील एकभाषिक प्रवाह आणि डिजिटल माध्यमांतील मर्यादित प्रतिनिधित्व यांमुळे अनेक भारतीय भाषा संकटात आहेत. डिजिटल युगात अस्तित्व टिकवण्यासाठी भाषांना केवळ सांस्कृतिक आधार नव्हे तर तांत्रिक अधिष्ठान आवश्यक आहे. नैसर्गिक भाषा प्रक्रिया (Natural Language Processing – NLP) ही कृत्रिम बुद्धिमत्तेची एक महत्त्वपूर्ण शाखा असून संगणकाला मानवी भाषा समजणे, विश्लेषित करणे आणि प्रतिसाद देणे शक्य करते. NLP च्या साहाय्याने भाषांचे डिजिटायझेशन, कॉर्पस निर्मिती, यांत्रिक भाषांतर, ध्वनी-ओळख (Speech Recognition), स्वयंचलित सारांश, भावना विश्लेषण (Sentiment Analysis) आणि बहुभाषिक संवाद साधनांची निर्मिती शक्य होते. भारतीय भाषांच्या संवर्धनाच्या दृष्टीने ही तंत्रप्रणाली क्रांतिकारक ठरू शकते. प्रस्तुत शोधनिबंधात भारतीय भाषांच्या डिजिटल संवर्धनासाठी NLP च्या उपयोगाचा आंतरविद्याशाखीय अभ्यास करण्यात आला आहे. भाषाशास्त्र, संगणकशास्त्र, डिजिटल मानवविद्या, सांस्कृतिक अध्ययन आणि सार्वजनिक धोरण या विविध शाखांचा समन्वय साधून एक एकात्मिक संवर्धन मॉडेल सुचविण्यात आले आहे. संशोधनात भारतीय भाषांसमोरील तांत्रिक, सामाजिक आणि धोरणात्मक आव्हाने विश्लेषित केली असून भविष्यातील दिशा प्रस्तावित केल्या आहेत.

मुख्य शब्द: नैसर्गिक भाषा प्रक्रिया (NLP), भारतीय भाषा, भाषा संवर्धन, डिजिटल मानवविद्या, कृत्रिम बुद्धिमत्ता, संगणकीय भाषाविज्ञान, मशीन ट्रान्सलेशन, कॉर्पस लिंग्विस्टिक्स, भाषिक वारसा, आंतरविद्याशाखीय संशोधन.

► *Corresponding Author: डॉ. अतुल मधुकरराव देशमुख*

प्रस्तावना

भाषा ही केवळ संवादाचे माध्यम नसून ती समाजाच्या ऐतिहासिक स्मृतीची, सांस्कृतिक ओळखीची आणि ज्ञानपरंपरेची वाहक असते. प्रत्येक भाषेमध्ये त्या समाजाचे तत्त्वज्ञान, लोककला, साहित्य, धार्मिक संकल्पना, सामाजिक रचना आणि मूल्यव्यवस्था प्रतिबिंबित होत असते. भारतासारख्या बहुभाषिक देशात भाषा ही सांस्कृतिक वैविध्याची पायाभूत रचना आहे.

भारतामध्ये २२ अनुसूचित भाषा आणि सुमारे १२२ प्रमुख भाषा तसेच शेकडो बोलीभाषा अस्तित्वात आहेत. तथापि, अनेक आदिवासी आणि प्रादेशिक भाषा लुप्त होण्याच्या मार्गावर आहेत. जागतिकीकरणामुळे इंग्रजी आणि काही प्रमुख भाषांचे वर्चस्व वाढले असून लहान भाषांची उपेक्षा होत आहे. शिक्षण, प्रशासन, तंत्रज्ञान आणि माध्यमे या क्षेत्रांत बहुधा काही मोजक्या भाषाच प्रचलित आहेत.

डिजिटल युगात ज्ञाननिर्मितीचे स्वरूप पूर्णतः बदलले आहे. इंटरनेट, स्मार्टफोन, कृत्रिम बुद्धिमत्ता आणि क्लाउड तंत्रज्ञानामुळे माहितीचे संग्रहण, प्रसारण आणि विश्लेषण व्यापक प्रमाणावर शक्य झाले आहे. परंतु डिजिटल क्षेत्रात भारतीय भाषांचे प्रतिनिधित्व मर्यादित आहे. अनेक भाषांसाठी पर्याप्त डिजिटल कॉर्पस उपलब्ध नाही; तसेच तांत्रिक साधने इंग्रजीकेंद्री विकसित झालेली आहेत.

या पार्श्वभूमीवर नैसर्गिक भाषा प्रक्रिया (NLP) ही तंत्रज्ञान शाखा भारतीय भाषांच्या संवर्धनासाठी प्रभावी ठरू शकते. NLP च्या साहाय्याने-
भाषांचे डिजिटल दस्तऐवजीकरण
बहुभाषिक यांत्रिक भाषांतर
भाषिक विश्लेषण
मातृभाषेतील शिक्षण साधने
सांस्कृतिक वारसा जतन
ही उद्दिष्टे साध्य करता येतात.
या शोधनिबंधाचा मुख्य हेतू म्हणजे भारतीय भाषांच्या डिजिटल संवर्धनासाठी NLP चे आंतरविद्याशाखीय योगदान तपासणे.

संशोधन समस्या

भारतीय भाषांच्या डिजिटल अस्तित्वासमोरील प्रमुख समस्या पुढीलप्रमाणे आहेत.
१. अनेक भाषांसाठी पुरेशा प्रमाणात डिजिटाइज्ड मजकूर उपलब्ध नाही.
२. लो-रिसोर्स भाषांसाठी NLP मॉडेल विकसित करणे कठीण जाते.
३. लिपीभेद आणि रूपवैविध्य (Morphological Richness) मोठे आहे.
४. कोड-मिश्रण (उदा. हिंग्लिश) यामुळे भाषिक विश्लेषण गुंतागुंतीचे होते.
५. भाषिक डेटा मानकीकरणाचा अभाव आहे.
या पार्श्वभूमीवर प्रश्न निर्माण होतो की NLP च्या साहाय्याने भारतीय भाषांचे संवर्धन प्रभावीपणे कसे करता येईल?

संशोधन प्रश्न

१. भारतीय भाषांच्या संवर्धनासाठी NLP चे कोणते तांत्रिक घटक सर्वाधिक उपयुक्त आहेत?
२. भारतीय भाषांतील रूपवैविध्य आणि लिपीभेद NLP साठी कोणती आव्हाने निर्माण करतात?
३. लो-रिसोर्स भाषांसाठी Transfer Learning आणि Multilingual Models किती प्रभावी ठरू शकतात?
४. डिजिटल मानवविद्या आणि NLP यांचा संगम सांस्कृतिक वारसा जतनासाठी कसा उपयुक्त आहे?
५. राष्ट्रीय स्तरावर भाषा संवर्धनासाठी कोणते धोरणात्मक मॉडेल विकसित करता येईल?

संशोधनाची उद्दिष्टे

१. भारतीय भाषांच्या डिजिटल स्थितीचे विश्लेषण करणे.
२. NLP तंत्रज्ञानाची उपयुक्तता आणि मर्यादा तपासणे.
३. भारतीय भाषांसाठी आंतरविद्याशाखीय संवर्धन मॉडेल सुचविणे.
४. धोरणात्मक आणि शैक्षणिक शिफारसी मांडणे.
५. भावी संशोधनाच्या दिशा स्पष्ट करणे.

अभ्यासाची व्याप्ती

हा अभ्यास मुख्यतः पुढील घटकांवर केंद्रित आहे.
भारतीय अनुसूचित व प्रादेशिक भाषा
डिजिटल दस्तऐवजीकरण
मशीन ट्रान्सलेशन
स्पीच प्रोसेसिंग
डिजिटल मानवविद्या
लो-रिसोर्स भाषांचे तांत्रिक उपाय

अभ्यासाची मर्यादा

सर्व भारतीय भाषांचा सविस्तर तांत्रिक विश्लेषण या अभ्यासात समाविष्ट नाही.
प्रत्यक्ष प्रायोगिक (Experimental) मॉडेल चाचणी करण्यात आलेली नाही.
सांकेतिक व संकल्पनात्मक पातळीवर विश्लेषण केले आहे.

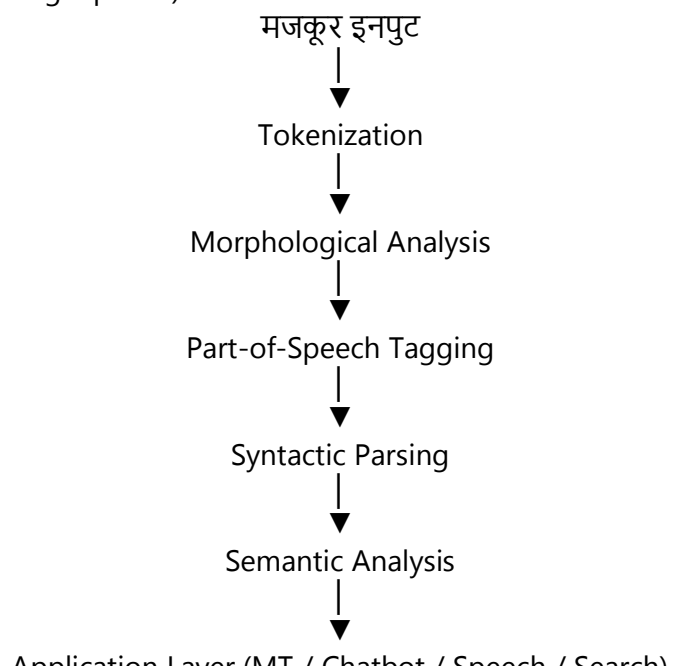
साहित्याचा सविस्तर आढावा

जागतिक पातळीवर NLP संशोधनात प्रचंड प्रगती झाली आहे. विशेषतः इंग्रजी भाषेसाठी Transformer आधारित मॉडेल्स (उदा. BERT, GPT, T5) विकसित झाली आहेत. बहुभाषिक मॉडेल्समुळे विविध भाषांसाठी Transfer Learning शक्य झाले आहे. तथापि, भारतीय भाषांच्या संदर्भात संशोधन अजून विकसित होत आहे.
कॉर्पस लिंग्विस्टिक्स, मशीन ट्रान्सलेशन आणि स्पीच प्रोसेसिंग या क्षेत्रात काही शासकीय आणि शैक्षणिक उपक्रम राबविण्यात आले आहेत. परंतु एकात्मिक राष्ट्रीय स्तरावरील समन्वित धोरणाचा अभाव दिसून येतो.
डिजिटल मानवविद्या या क्षेत्रात साहित्य, इतिहास, लोककथा, सांस्कृतिक अभिलेख यांचे संगणकीय विश्लेषण केले जाते. NLP आणि डिजिटल मानवविद्या यांच्या संगमातून भारतीय भाषांच्या संवर्धनासाठी व्यापक संधी उपलब्ध होऊ शकतात.

नैसर्गिक भाषा प्रक्रिया : तांत्रिक पाया

नैसर्गिक भाषा प्रक्रिया (NLP) ही कृत्रिम बुद्धिमत्तेची उपशाखा असून संगणकाला मानवी भाषा समजून घेणे, विश्लेषण करणे आणि प्रतिसाद देणे शक्य करते. NLP चे कार्य मुख्यतः दोन स्तरांवर विभागले जाते:
१. नियमाधारित (Rule-based systems)
२. सांख्यिकीय व मशीन लर्निंग आधारित प्रणाली (Statistical & ML-based systems)
अलीकडील काळात डीप लर्निंग (Deep Learning) आणि Transformer आधारित मॉडेल्स (उदा. BERT, GPT, mT5) यांनी NLP मध्ये आमूलाग्र बदल घडवून आणला आहे.

NLP प्रक्रिया स्तर (Processing Pipeline)



भारतीय भाषांसाठी या प्रत्येक स्तरावर स्वतंत्र संशोधन आवश्यक आहे.

भारतीय भाषांची भाषिक वैशिष्ट्ये आणि NLP आव्हाने

भारतीय भाषा प्रामुख्याने दोन प्रमुख भाषा परिवारांत मोडतात:
इंडो-आर्यन (उदा. हिंदी, मराठी, बंगाली)
द्रविड (उदा. तमिळ, तेलुगू, कन्नड, मल्याळम)

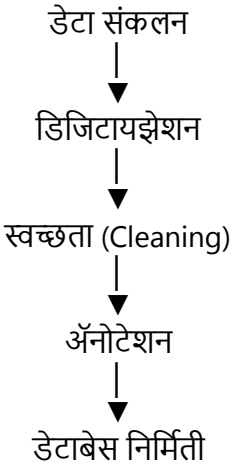
रचनात्मक वैशिष्ट्ये
घटक
वैशिष्ट्य
शब्दरचना
रूपवैविध्यपूर्ण (Morphologically rich)
वाक्यरचना
SOV (Subject-Object-Verb)
लिपी
बहुलिपीय (देवनागरी, गुरुमुखी, तमिळ इ.)
संयुक्त शब्द
विपुल प्रमाणात
Sandhi व समास
वारंवार आढळतात
NLP साठी आव्हाने:
एकाच शब्दाचे अनेक रूप
विभक्ती प्रत्ययांची जटिलता
लिंग, वचन, पुरुष भिन्नता
बोलीभाषा फरक
कोड-मिश्रण (उदा. "मी meeting ला जात आहे")

कॉर्पस निर्मिती आणि डेटा व्यवस्थापन

NLP प्रणालीसाठी मोठ्या प्रमाणात डेटा (Corpus) आवश्यक असतो.

कॉर्पसचे प्रकार
प्रकार
उदाहरण
लिखित कॉर्पस
साहित्य, वृत्तपत्रे
मौखिक कॉर्पस
लोककथा, मुलाखती
समांतर कॉर्पस
भाषांतरित मजकूर
ॲनोटेटेड कॉर्पस
POS, NER टॅगिंग

कॉर्पस निर्मिती प्रक्रिया



भारतीय भाषांसाठी मोठ्या प्रमाणावर मुक्त स्रोत कॉर्पसची आवश्यकता आहे.

मशीन ट्रान्सलेशन

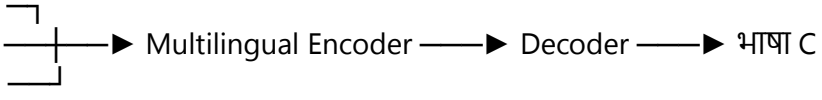
भारतीय भाषांमधील ज्ञान आदानप्रदानासाठी मशीन ट्रान्सलेशन अत्यंत महत्त्वाचे आहे.

प्रकार

१. Rule-Based MT
२. Statistical MT
३. Neural Machine Translation (NMT)

Neural MT मध्ये Transformer आधारित मॉडेल्स प्रभावी ठरतात.

बहुभाषिक मॉडेल प्रवाह



भारतीय संदर्भातील उपयोग

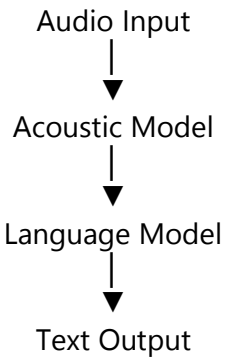
प्रशासनिक कागदपत्रांचे भाषांतर
शैक्षणिक सामग्रीचे स्थानिकीकरण
न्यायालयीन भाषांतर सहाय्य

स्पीच प्रोसेसिंग

भारतासारख्या देशात मौखिक परंपरा अत्यंत मजबूत आहे. त्यामुळे स्पीच-आधारित NLP साधने महत्त्वाची ठरतात.

प्रमुख घटक

Automatic Speech Recognition (ASR)
Text-to-Speech (TTS)
Voice Assistants
स्पीच प्रोसेसिंग मॉडेल



आव्हाने:
बोलीभाषा विविधता
उच्चार भिन्नता
पार्श्वध्वनी (Noise)

डिजिटल मानवविद्या आणि NLP

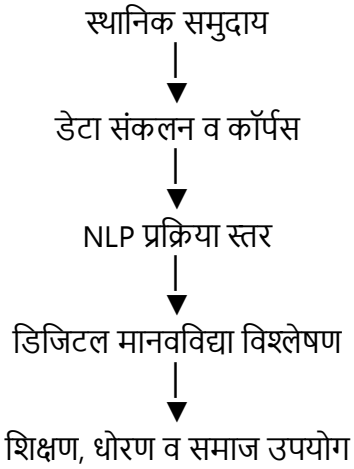
डिजिटल मानवविद्या (Digital Humanities) या क्षेत्रात साहित्य, इतिहास, सांस्कृतिक अभ्यास यांचे संगणकीय विश्लेषण केले जाते.

अनुप्रयोग
क्षेत्र
NLP उपयोग
संत साहित्य
शब्द वारंवारता विश्लेषण
लोककथा
थीम मॉडेलिंग
ऐतिहासिक दस्तऐवज
नामनिर्देशन (NER)
वृत्तपत्रे
भावना विश्लेषण

लो-रिसोर्स भाषांसाठी उपाय

भारतीय भाषांपैकी अनेक भाषा लो-रिसोर्स आहेत.
उपाय:
१. Transfer Learning
२. Multilingual Pre-trained Models
३. Crowdsourcing Data
४. Open-source Platforms

भारतीय भाषांसाठी एकात्मिक संवर्धन मॉडेल



हे मॉडेल तांत्रिक आणि सांस्कृतिक घटकांचा समन्वय साधते.

चर्चा

Transformer आधारित बहुभाषिक मॉडेल्समुळे भारतीय भाषांसाठी समान पायाभूत तंत्र विकसित करणे शक्य झाले आहे. तथापि, डेटाचा अभाव आणि ॲनोटेशन खर्च ही मोठी अडचण आहे. शासन, शैक्षणिक संस्था आणि तंत्रज्ञान उद्योग यांचा समन्वय आवश्यक आहे.

सामाजिक व सांस्कृतिक परिणाम

भारतीय भाषांचे संवर्धन हे केवळ तांत्रिक कार्य नसून सामाजिक, सांस्कृतिक आणि राजकीय दृष्टिकोनातूनही महत्त्वाचे आहे. NLP च्या माध्यमातून भाषांचे डिजिटायझेशन केल्यास पुढील परिणाम दिसून येतात:

सांस्कृतिक ओळख बळकटीकरण

भाषा ही ओळखीचा मुख्य घटक आहे. डिजिटल माध्यमात भाषा दृश्यमान झाल्यास तरुण पिढीला तिच्याशी जोडले जाणे सोपे होते.

ज्ञानलोकशाही

मातृभाषेत डिजिटल साधने उपलब्ध झाल्यास ज्ञान सर्व स्तरांपर्यंत पोहोचते. ग्रामीण व अल्पशिक्षित घटकांनाही माहिती सहज मिळते.

अल्पसंख्याक भाषांचे संरक्षण

NLP आधारित दस्तऐवजीकरणामुळे लुप्तप्राय बोली जतन करता येतात.

नैतिकता आणि जबाबदाऱ्या

भारतीय भाषांसाठी NLP विकसित करताना काही नैतिक मुद्दे विचारात घ्यावे लागतात:

१. डेटा गोपनीयता

२. भाषिक प्रतिनिधित्वातील पक्षपात (Bias)

३. अल्पसंख्याक समुदायांचा सहभाग

४. व्यावसायिक शोषण टाळणे

AI मॉडेल्समध्ये भाषिक पक्षपात राहू नये यासाठी विविध स्रोतांमधील संतुलित डेटा आवश्यक आहे.

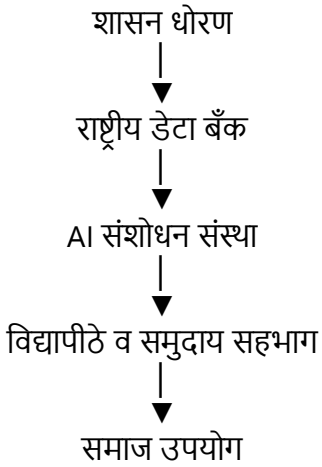
धोरणात्मक आराखडा

राष्ट्रीय स्तरावरील उपक्रम

राष्ट्रीय भाषिक कॉर्पस निर्मिती

मुक्त स्रोत NLP प्लॅटफॉर्म
विद्यापीठ-उद्योग सहकार्य
डिजिटल संग्रहालये

प्रस्तावित धोरण मॉडेल



शिक्षण आणि संशोधनातील भूमिका
भारतीय भाषांमध्ये डिजिटल शिक्षण साधने विकसित केल्यास:
मातृभाषा शिक्षण सुलभ
संशोधनासाठी कॉर्पस उपलब्ध
ग्रामीण विद्यार्थ्यांसाठी डिजिटल सहाय्य
उच्च शिक्षण संस्थांमध्ये "Digital Humanities & NLP" या विषयाचा अभ्यासक्रम सुरू करणे आवश्यक आहे.

भावी दिशा
१. बहुभाषिक Large Language Models विकसित करणे
२. लो-रिसोर्स भाषांसाठी विशेष संशोधन निधी
३. भारतीय लिप्यांसाठी मानकीकरण
४. स्पीच-आधारित ग्रामीण सहाय्यक प्रणाली
५. भारतीय भाषांसाठी ओपन डेटा चळवळ

व्यापक चर्चा
भारतीय भाषांचे संवर्धन करण्यासाठी केवळ तांत्रिक साधने पुरेशी नाहीत. सामाजिक सहभाग, शैक्षणिक धोरण आणि आर्थिक पाठबळ आवश्यक आहे. Transformer आणि Neural Networks सारख्या आधुनिक तंत्रज्ञानामुळे बहुभाषिक मॉडेल्स विकसित करणे शक्य झाले आहे; परंतु भाषिक विविधता लक्षात घेऊन स्थानिक पातळीवर संशोधन करणे आवश्यक आहे.
आंतरविद्याशाखीय दृष्टिकोनातून पाहता—
भाषाशास्त्र रचनात्मक विश्लेषण देते
संगणकशास्त्र तांत्रिक उपाय देते
मानवविद्या सांस्कृतिक संदर्भ देते
सार्वजनिक धोरण अंमलबजावणी सुनिश्चित करते
या सर्वांचा समन्वय भारतीय भाषांचे दीर्घकालीन संरक्षण सुनिश्चित करू शकतो.

निष्कर्ष

भारतीय भाषांचे संवर्धन हे सांस्कृतिक वारशाचे रक्षण करण्याइतकेच ज्ञानव्यवस्थेचे लोकशाहीकरण करण्यासाठीही अत्यावश्यक आहे. नैसर्गिक भाषा प्रक्रिया (NLP) ही तंत्रज्ञान शाखा भारतीय भाषांना डिजिटल युगात सशक्त स्थान देऊ शकते. कॉर्पस निर्मिती, मशीन ट्रान्सलेशन, स्पीच प्रोसेसिंग, डिजिटल मानवविद्या आणि बहुभाषिक AI मॉडेल्स यांच्या साहाय्याने भाषांचे दस्तऐवजीकरण, विश्लेषण आणि प्रसार शक्य होतो.

भविष्यात शासन, शैक्षणिक संस्था, तंत्रज्ञान उद्योग आणि स्थानिक समुदाय यांच्या समन्वयातून भारतीय भाषांसाठी एकात्मिक डिजिटल पर्यावरण निर्माण करणे आवश्यक आहे. हा शोधनिबंध भारतीय भाषांच्या संवर्धनासाठी एक सैद्धांतिक व धोरणात्मक चौकट प्रदान करतो.

संदर्भसूची

1. Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology, 6(3), 1–26.
2. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT.
4. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Draft.
5. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
6. Rao, D., & Kulkarni, P. (2019). Natural language processing for Indian languages: Challenges and perspectives. Journal of Language Technology, 5(2), 45–58.
7. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
8. UNESCO. (2019). Atlas of the World's Languages in Danger. UNESCO Publishing.