

भारतीय भाषिक वारसा जतन करण्यासाठी नैसर्गिक भाषा प्रक्रिया (NLP): संधी, आव्हाने आणि उपाययोजना

प्रा. आत्माराम तुकाराम चिमकर

Assistant Professor, मराठी विभाग प्रमुख, दादासाहेब दि. शं. पाटील कला, वाणिज्य व विज्ञान महाविद्यालय,
एरंडोल, ता. एरंडोल जि. जळगाव पिन: ४२५१०९.

Email: atuchimkar@gmail.com

सारांश:

भारतीय भाषिक वारसा अत्यंत समृद्ध, विविध आणि सांस्कृतिकदृष्ट्या महत्त्वपूर्ण आहे. भारतात शेकडो भाषा आणि बोलीभाषा प्रचलित असून त्या देशाच्या ऐतिहासिक, सामाजिक आणि सांस्कृतिक ओळखीचा आधार आहेत. तथापि, जागतिकीकरण, डिजिटल वर्चस्व आणि इंग्रजीच्या वाढत्या प्रभावामुळे अनेक स्थानिक भाषा लुप्त होण्याच्या धोक्यात आहेत. या पार्श्वभूमीवर नैसर्गिक भाषा प्रक्रिया तंत्रज्ञान भारतीय भाषांचे जतन, संवर्धन आणि प्रसार करण्यासाठी महत्त्वपूर्ण ठरू शकते. भाषांतर साधने, वाणी-ओळख प्रणाली, मजकूर विश्लेषण, डिजिटल संग्रहण आणि शैक्षणिक अनुप्रयोगांद्वारे भाषिक वारसा डिजिटल स्वरूपात सुरक्षित ठेवता येतो. तरीही, अपुरी डेटासंचय, भाषिक विविधता, तांत्रिक अडचणी आणि धोरणात्मक कमतरता ही मोठी आव्हाने आहेत. योग्य धोरणे, संशोधन, सरकारी पाठबळ आणि तांत्रिक नवकल्पना यांच्या साहाय्याने नैसर्गिक भाषा प्रक्रिया(NLP)भारतीय भाषांच्या टिकावासाठी प्रभावी साधन ठरू शकते.

मुख्य शब्द: नैसर्गिक भाषा प्रक्रिया (NLP), भारतीय भाषिक वारसा, डिजिटल संवर्धन, भाषांतर तंत्रज्ञान, वाणी-ओळख प्रणाली, भाषिक विविधता, कृत्रिम बुद्धिमत्ता व भाषिक धोरण.

► Corresponding Author: प्रा. आत्माराम तुकाराम चिमकर

प्रस्तावना:

भारत हा जगातील सर्वात समृद्ध बहुभाषिक देशांपैकी एक असून येथे २२ अनुसूचित भाषा आणि शेकडो प्रादेशिक बोलीभाषा प्रचलित आहेत. प्रत्येक भाषा ही त्या प्रदेशाच्या इतिहास, परंपरा, साहित्य आणि सांस्कृतिक ओळखीचे प्रतीक आहे. भारतीय भाषांमधील विविधता देशाच्या सामाजिक ऐक्याला आणि सांस्कृतिक वैभवाला बळकटी देते. तथापि, "जागतिकीकरण, शहरीकरण आणि इंग्रजीच्या वाढत्या प्रभावामुळे स्थानिक भाषांचा वापर डिजिटल आणि शैक्षणिक क्षेत्रात तुलनेने कमी होत आहे. अनेक लहान बोलीभाषा लुप्त होण्याच्या धोक्यात आहेत.या पार्श्वभूमीवर नैसर्गिक भाषा प्रक्रिया (Natural Language Processing – NLP) ही कृत्रिम बुद्धिमत्तेची महत्त्वाची शाखा भाषेचे संगणकीय विश्लेषण, प्रक्रिया आणि रूपांतरण करण्यास सक्षम आहे."^१ भाषांतर, वाणी-ओळख, मजकूर विश्लेषण आणि डिजिटल संग्रहण यांसारख्या तंत्रज्ञानाच्या साहाय्याने भारतीय भाषांचे जतन आणि प्रसार अधिक प्रभावीपणे करता येऊ शकतो. त्यामुळे नैसर्गिक भाषा प्रक्रिया(NLP)हे भारतीय भाषिक वारसा संवर्धनासाठी एक महत्त्वपूर्ण आणि भविष्याभिमुख साधन ठरू शकते.

१. भाषांतर तंत्रज्ञान:

भाषांतर तंत्रज्ञान हे नैसर्गिक भाषा प्रक्रिया (NLP) मधील अत्यंत महत्त्वाचे क्षेत्र आहे. "डिजिटल युगात विविध भाषांमधील संवाद सुलभ करण्यासाठी स्वयंचलित भाषांतर साधनांची भूमिका मोठी ठरली आहे.गुगल ट्रान्सलेट सारख्या प्लॅटफॉर्ममुळे मराठी, हिंदी, तमिळ, बंगाली तसेच इतर अनेक भारतीय भाषांमधील मजकूराचे तत्काळ भाषांतर शक्य झाले आहे."^२ यामुळे ज्ञान, माहिती, शैक्षणिक सामग्री आणि शासकीय सेवा विविध भाषिक समुदायांपर्यंत पोहोचू शकतात.भाषांतर तंत्रज्ञानामुळे जागतिक माहिती स्थानिक भाषांमध्ये उपलब्ध होत असल्याने डिजिटल समावेशन

वाढते. "ग्रामीण आणि अल्पशिक्षित नागरिकांसाठीही ऑनलाइन सेवा, शासकीय योजना आणि शैक्षणिक साधने अधिक सुलभ होतात. तसेच व्यवसाय, पर्यटन आणि आंतरराष्ट्रीय संवादामध्येही याचा मोठा फायदा होतो." तथापि, भाषांतर प्रणालींमध्ये अजूनही संदर्भ, वाक्प्रचार, सांस्कृतिक सूक्ष्मता आणि बोलीभाषेतील फरक यांचे अचूक रूपांतर करणे हे आव्हान आहे. त्यामुळे मानवी पुनरावलोकन आणि स्थानिक भाषातज्ज्ञांचा सहभाग आवश्यक ठरतो. एकूणच, भाषांतर तंत्रज्ञान भारतीय भाषांच्या जतन, प्रसार आणि डिजिटल सक्षमीकरणासाठी प्रभावी साधन ठरत आहे.

२. वाणीओळख प्रणाली:

"वाणीओळख प्रणाली (Speech Recognition System) ही नैसर्गिक भाषा प्रक्रिया (NLP) मधील एक महत्त्वाची तांत्रिक प्रगती आहे. या तंत्रज्ञानाद्वारे संगणक किंवा मोबाईल उपकरणे मानवी आवाज ओळखून त्याचे मजकूर रूपांतर करू शकतात किंवा आवाजाच्या आदेशानुसार कार्य करू शकतात." मायक्रोसॉफ्ट आणिगुगल यांनी मराठी, हिंदी, तमिळ, तेलुगू यांसारख्या अनेक भारतीय भाषांसाठी वाणी-ओळख प्रणाली विकसित केल्या आहेत. या प्रणालींचा सर्वात मोठा फायदा ग्रामीण आणि अल्पशिक्षित नागरिकांना होतो. ज्यांना टायपिंग अवघड जाते किंवा इंग्रजी भाषेचे ज्ञान मर्यादित आहे, ते आपल्या मातृभाषेत आवाजाद्वारे डिजिटल सेवा वापरू शकतात. उदाहरणार्थ, शासकीय योजना, बँकिंग सेवा, आरोग्यविषयक माहिती किंवा शैक्षणिक अॅप्समध्ये आवाजाच्या माध्यमातून संवाद साधणे शक्य होते. तथापि, भारतीय भाषांतील उच्चारांतील विविधता, बोलीभाषा आणि पार्श्वभूमीतील आवाज ही मोठी आव्हाने आहेत. त्यामुळे अचूकता वाढवण्यासाठी मोठ्या प्रमाणात स्थानिक डेटा संकलन आणि सातत्याने तांत्रिक सुधारणा आवश्यक आहेत. एकूणच, वाणी-ओळख प्रणाली डिजिटल समावेशन वाढवून भारतीय भाषांच्या संवर्धनाला चालना देतात.

३. डिजिटल संग्रहण आणि अभिलेखन:

डिजिटल संग्रहण आणि अभिलेखन ही भारतीय भाषिक वारसा जतन करण्यासाठी अत्यंत महत्त्वाची प्रक्रिया आहे. भारतातील प्राचीन साहित्य, संतवाङ्मय, लोकगीते, ताडपत्रीवरील हस्तलिखिते आणि दुर्मिळ ग्रंथ हे सांस्कृतिक संपत्तीचे अमूल्य स्रोत आहेत. कालांतराने कागद, शाई आणि भौतिक माध्यमांची झीज होत असल्यामुळे या वारशाचे जतन करणे आवश्यक ठरते. यासाठी नैसर्गिक भाषा प्रक्रिया (NLP) आधारित टेक्स्ट डिजिटायझेशन साधने मोठी मदत करतात. "ऑप्टिकल कॅरेक्टर रिकग्निशन (OCR) तंत्रज्ञानाद्वारे स्कॅन केलेल्या दस्तऐवजांमधील मजकूर संगणकीय स्वरूपात रूपांतरित केला जातो. त्यानंतर नैसर्गिक भाषा प्रक्रिया(NLP)साधनांच्या मदतीने मजकूर दुरुस्ती, भाषिक विश्लेषण, अनुक्रमणिका तयार करणे आणि शोधसुलभता वाढवणे शक्य होते." यामुळे संशोधक, विद्यार्थी आणि सर्वसामान्य नागरिकांना दुर्मिळ साहित्य ऑनलाइन उपलब्ध होते. तथापि, विविध लिपी, जुनी शब्दरचना आणि हस्ताक्षरातील वैविध्य यामुळे अचूक डिजिटायझेशन करणे आव्हानात्मक ठरते. त्यामुळे स्थानिक भाषातज्ज्ञ आणि तांत्रिक तज्ञ यांचे सहकार्य आवश्यक आहे. एकूणच, डिजिटल संग्रहणामुळे भारतीय भाषिक वारसा सुरक्षित राहून पुढील पिढ्यांपर्यंत पोहोचवता येतो.

४. शैक्षणिक उपयोग:

नैसर्गिक भाषा प्रक्रिया (NLP) तंत्रज्ञानाचा शैक्षणिक क्षेत्रात अत्यंत व्यापक आणि सकारात्मक उपयोग होत आहे. "ई-लर्निंग प्लॅटफॉर्मवर स्थानिक भाषांमध्ये शैक्षणिक सामग्री उपलब्ध करून विद्यार्थ्यांना शिक्षण अधिक सुलभ आणि समजण्यास सोपे केले जाते. भारतासारख्या बहुभाषिक देशात अनेक विद्यार्थ्यांना इंग्रजीपेक्षा मातृभाषेत शिक्षण घेतल्यास विषय अधिक चांगल्या प्रकारे समजते." त्यामुळे डिजिटल शिक्षण साधनांमध्ये मराठी, हिंदी, तमिळ, बंगाली इत्यादी भाषांतील सामग्री उपलब्ध करणे अत्यंत आवश्यक ठरते.

"नैसर्गिक भाषा प्रक्रिया(NLP)आधारित भाषांतर, वाणी-ओळख आणि टेक्स्ट-टू-स्पीच तंत्रज्ञानामुळे विद्यार्थ्यांना अभ्यास साहित्य ऐकणे, वाचणे आणि समजून घेणे अधिक सोपे होते. उदाहरणार्थ, गुगल क्लासरूम आणिस्वयम (SWAYAM)सारख्या प्लॅटफॉर्मवर विविध भारतीय भाषांमध्ये अभ्यासक्रम उपलब्ध करून दिले जातात." यामुळे ग्रामीण भागातील आणि अल्पसंसाधन असलेल्या विद्यार्थ्यांनाही गुणवत्तापूर्ण शिक्षणाची संधी मिळते. याशिवाय, स्वयंचलित मूल्यमापन प्रणाली, चॅटबॉट्स आणि वैयक्तिकृत शिक्षण यांसारख्या सुविधा विद्यार्थ्यांच्या प्रगतीचे विश्लेषण करून त्यानुसार मार्गदर्शन करतात. तथापि, सर्व भाषांमध्ये दर्जेदार डिजिटल सामग्रीची कमतरता आणि तांत्रिक पायाभूत सुविधांची मर्यादा ही आव्हाने आहेत. तरीही, नैसर्गिक भाषा प्रक्रिया(NLP)च्या साहाय्याने स्थानिक भाषांमध्ये शिक्षण उपलब्ध करून शैक्षणिक समावेशन आणि समान संधी निर्माण करता येतात.

आव्हाने:

भारतीय भाषिक वारसा जतन करण्यासाठी नैसर्गिक भाषा प्रक्रिया (NLP) अत्यंत उपयुक्त असली तरी त्यासमोर अनेक गंभीर आव्हाने उभी आहेत.

१. सर्वप्रथम, भारतीय भाषांमध्ये गुणवत्तापूर्ण आणि प्रमाणित माहिती संचयन करण्याची कमतरता ही मोठी समस्या आहे. नैसर्गिक भाषा प्रक्रिया(NLP)मॉडेल्स प्रभावीपणे कार्य करण्यासाठी मोठ्या प्रमाणावर स्वच्छ, विविध आणि अचूक भाषिक डेटा आवश्यक असतो. परंतु अनेक प्रादेशिक भाषा आणि बोलीभाषांसाठी पुरेसा डिजिटल डेटा उपलब्ध नाही. त्यामुळे या भाषांसाठी अचूक भाषांतर, वाणी-ओळख किंवा मजकूर विश्लेषण प्रणाली विकसित करणे कठीण ठरते.
२. दुसरे महत्त्वाचे आव्हान म्हणजे विविध लिपींचे अस्तित्व. देवनागरी, तमिळ, तेलुगू, गुरुमुखी, बंगाली अशा अनेक लिपींमुळे तांत्रिक प्रक्रिया अधिक गुंतागुंतीची होते. काही भाषांमध्ये एकाच शब्दाचे अनेक उच्चार आणि लेखनभेद आढळतात, ज्यामुळे संगणकीय विश्लेषणात अडचणी येतात.
३. तिसरे आव्हान म्हणजे व्याकरणातील आणि वाक्यरचनेतील गुंतागुंत. भारतीय भाषांमध्ये लिंग, वचन, विभक्ती, काळ आणि शब्दक्रम यांचे विविध प्रकार आढळतात. बोलीभाषा आणि प्रादेशिक फरक यामुळेही अचूक मॉडेल तयार करणे अवघड होते.
४. याशिवाय, तांत्रिक पायाभूत सुविधांची कमतरता, संशोधनासाठी निधीअभाव आणि प्रशिक्षित तज्ज्ञांची कमी संख्या हीदेखील महत्त्वाची अडचण आहे. त्यामुळे भारतीय भाषांसाठी प्रभावी नैसर्गिक भाषा प्रक्रिया(NLP)उपाय विकसित करण्यासाठी व्यापक संशोधन, सहकार्य आणि धोरणात्मक पाठबळ आवश्यक आहे.

उपाययोजना:

भारतीय भाषिक वारसा जतन करण्यासाठी नैसर्गिक भाषा प्रक्रिया (NLP) प्रभावीपणे वापरायची असेल तर सर्वसमावेशक आणि दीर्घकालीन उपाययोजना आवश्यक आहेत.

१. सर्वप्रथम, केंद्र व राज्य सरकारांनी स्पष्ट आणि ठोस भाषिक धोरणे आखून भारतीय भाषांच्या डिजिटल संवर्धनाला प्राधान्य द्यावे. शैक्षणिक संस्था, तंत्रज्ञान विद्यापीठे आणि संशोधन केंद्रांना पुरेसा निधी व पायाभूत सुविधा उपलब्ध करून दिल्यास स्थानिक भाषांसाठी दर्जेदार नैसर्गिक भाषा प्रक्रिया(NLP)साधने विकसित करता येतील.
२. दुसरे महत्त्वाचे पाऊल म्हणजे मुक्त-स्रोत (Open Source) डेटासंचय तयार करणे. विविध भारतीय भाषांतील साहित्य, संभाषणे, वृत्तपत्रे आणि शैक्षणिक मजकूर यांचे डिजिटल संकलन करून संशोधकांसाठी खुले करणे आवश्यक आहे. यामुळे भाषिक मॉडेल्स अधिक अचूक आणि सर्वसमावेशक बनतील.
३. तिसरे म्हणजे खासगी क्षेत्राचा सक्रिय सहभाग. तंत्रज्ञान कंपन्यांनी स्थानिक भाषांसाठी वाणी-ओळख, भाषांतर आणि टेक्स्ट-टू-स्पीच प्रणाली विकसित करण्यावर भर द्यावा. यासाठी स्थानिक भाषातज्ज्ञ, व्याकरण तज्ञ आणि सांस्कृतिक अभ्यासक यांचा सहभाग अत्यंत महत्त्वाचा आहे.
४. याशिवाय, सार्वजनिक-खासगी भागीदारी (PPP) मॉडेलद्वारे संशोधन आणि अंमलबजावणीला गती देता येईल. प्रशिक्षण कार्यक्रमांद्वारे युवकांना भाषिक तंत्रज्ञानात कौशल्य विकसित करण्याची संधी दिल्यास रोजगारनिर्मितीही होईल. अशा समन्वित प्रयत्नांमुळे भारतीय भाषांचे डिजिटल जतन आणि प्रसार प्रभावीपणे साध्य होऊ शकते.

निष्कर्ष:

भारतीय भाषिक वारसा जतन करणे हे केवळ सांस्कृतिक कर्तव्य नसून सामाजिक, शैक्षणिक आणि राजकीय दृष्टीनेही अत्यंत महत्त्वाचे आहे. भाषा ही केवळ संवादाचे माध्यम नसून ती एखाद्या समाजाची ओळख, इतिहास, परंपरा आणि मूल्यव्यवस्था प्रतिबिंबित करते. जागतिकीकरण आणि डिजिटल माध्यमांच्या वाढत्या प्रभावामुळे अनेक स्थानिक भाषा आणि बोलीभाषा लुप्त होण्याच्या धोक्यात आहेत. अशा परिस्थितीत डिजिटल तंत्रज्ञानाचा प्रभावी वापर करणे अपरिहार्य ठरते.

नैसर्गिक भाषा प्रक्रिया (NLP) हे तंत्रज्ञान भारतीय भाषांच्या जतन आणि संवर्धनासाठी मोठी संधी उपलब्ध करून देते. भाषांतर प्रणालींमुळे विविध भाषांतील माहिती सहज उपलब्ध होते, वाणी-ओळख तंत्रज्ञानामुळे डिजिटल सेवा सर्वसामान्यांपर्यंत पोहोचतात, तर डिजिटल संग्रहणामुळे प्राचीन साहित्य आणि सांस्कृतिक दस्तऐवज सुरक्षित राहतात. तसेच शैक्षणिक क्षेत्रात स्थानिक भाषांमध्ये अभ्यास सामग्री उपलब्ध करून विद्यार्थ्यांना समावेशक शिक्षणाची संधी मिळते. तथापि, या प्रक्रियेत तांत्रिक मर्यादा, गुणवत्तापूर्ण डेटासंचयाची कमतरता, विविध लिपी आणि बोलीभाषांतील

गुंतागुंत ही मोठी आव्हाने आहेत. त्यामुळे सर्वकष आणि समन्वित धोरणांची गरज आहे. सरकार, संशोधन संस्था, तंत्रज्ञान कंपन्या आणि स्थानिक भाषातज्ज्ञ यांनी संयुक्तपणे कार्य केल्यास अधिक अचूक आणि सर्वसमावेशक नैसर्गिक भाषा प्रक्रिया(NLP)साधने विकसित होऊ शकतात.योग्य नियोजन, संशोधन आणि नवकल्पनांच्या सहाय्याने भारतीय भाषांना जागतिक डिजिटल विश्वात सशक्त स्थान मिळू शकते. यामुळे भावी पिढ्यांसाठी आपला भाषिक आणि सांस्कृतिक वारसा सुरक्षित राहिल आणि डिजिटल युगातही भारतीय भाषांची समृद्ध परंपरा टिकून राहिल.

संदर्भ:

१. जोशीहेमंत, डिजिटल मीडिया आणि भाषा, अर्णव प्रकाशन, पुणे, २०२४, पृ. क्र.
२. देशपांडेसतिश आणि कुलकर्णी, मिलिंद, संगणक आणि मराठी भाषा: नवे आयाम, कॉन्टिनेन्टल प्रकाशन, पुणे, २०२५, पृ. क्र.
३. शर्माराम अवतार, भारतीय भाषा विज्ञान एवं कंप्यूटेशनल लिंग्विस्टिक्स, वाणी प्रकाशन, नई दिल्ली, २०२३, पृ. क्र.
४. पटवर्धनअंजली, मराठी बोलीभाषांचे डिजिटलीकरण आणि आव्हाने, भाषा आणि जीवन, मराठी अभ्यास परिषद, पुणे, २०२४, पृ. क्र. ०५.
५. गायकवाड, विनायक, माहिती तंत्रज्ञान आणि मराठी साहित्याचा वारसा, मेहता पब्लिशिंग हाऊस, पुणे, २०२६, पृ. क्र.
६. www.bhashini.gov.in(भाषिणी - "नॅशनल लॅंग्वेज ट्रान्सलेशन मिशन: डिजिटल इंडिया उपक्रम", इलेक्ट्रॉनिक्स आणि माहिती तंत्रज्ञान मंत्रालय, भारत सरकार, २०२५.
७. राजभाषा विभाग, हिंदी और क्षेत्रीय भाषाओं का डिजिटल डेटाबेस, गृह मंत्रालय, भारत सरकार, २०२६, पृ. क्र.