

डिजिटल युग में हिंदी भाषा के बदलते स्वरूप का अध्ययन

डॉ. निर्मला परगनिहा

सहयक प्रोफेसर अग्रसेन कॉलेज

धनोरा, दुर्ग, छत्तीसगढ़



<http://www.ujhmds.org/> || Vol. 2 No. 3 (2026): Special August Issue

Date of Submission: 02-07-2026

Date of Acceptance: 22-07-2026

Date of Publication: 17-08-2026

सारांश— डिजिटल संचार, सोशल मीडिया, कृत्रिम बुद्धिमत्ता और मोबाइल प्रौद्योगिकी ने हिंदी भाषा के प्रयोग, संरचना और अभिव्यक्ति के तरीकों में उल्लेखनीय परिवर्तन उत्पन्न किए हैं। प्रस्तुत अध्ययन का उद्देश्य डिजिटल परिवेश में हिंदी के बदलते स्वरूप को शब्दावली, लिपि-प्रयोग, वाक्य-संरचना, संक्षिप्तीकरण तथा अंग्रेज़ी-हिंदी मिश्रण के संदर्भ में समझना है। अध्ययन विशेष रूप से इस शोध-अंतराल पर केंद्रित है कि डिजिटल हिंदी को प्रायः केवल ‘हिंग्लिश’ अथवा रोमन लिपि के बढ़ते प्रयोग तक सीमित करके देखा गया है, जबकि विभिन्न डिजिटल मंचों और कृत्रिम बुद्धिमत्ता आधारित संचार में होने वाले बहुस्तरीय भाषिक परिवर्तन का समेकित अध्ययन अपेक्षाकृत सीमित है। इस शोध में डिजिटल हिंदी को एक स्थिर भाषिक रूप न मानकर मंच, संचार-उद्देश्य और प्रौद्योगिकीय माध्यम के अनुसार परिवर्तित होने वाली गतिशील भाषिक व्यवस्था के रूप में परिकल्पित किया गया है। इसके अंतर्गत देवनागरी हिंदी, रोमनकृत हिंदी, हिंदी-अंग्रेज़ी कोड-मिश्रण, डिजिटल संक्षिप्तियों और कृत्रिम बुद्धिमत्ता द्वारा निर्मित हिंदी की भाषिक विशेषताओं को विश्लेषण की प्रमुख इकाइयों के रूप में प्रस्तावित किया गया है। अध्ययन की वैचारिक रूपरेखा भाषिक परिवर्तन, कंप्यूटर-मध्यस्थ संचार और प्राकृतिक भाषा संसाधन के अंतर्संबंध पर आधारित है। यह शोध डिजिटल प्रयोगों को हिंदी के क्षरण के रूप में देखने के बजाय उनके माध्यम से उत्पन्न नए भाषिक व्यवहार, अनुकूलन और मानकीकरण संबंधी चुनौतियों का संतुलित विश्लेषण प्रस्तुत करता है।

मुख्य शब्द— डिजिटल हिंदी, भाषिक परिवर्तन, हिंग्लिश, रोमनकृत हिंदी, कृत्रिम बुद्धिमत्ता, प्राकृतिक भाषा संसाधन

परिचय

भाषा सामाजिक परिवर्तन की एक संवेदनशील अभिव्यक्ति है। समाज में संचार के साधन, ज्ञान के स्रोत और पारस्परिक संपर्क के तरीके बदलने

पर भाषा भी नए परिवेश के अनुरूप अपने स्वरूप में परिवर्तन करती है। इक्कीसवीं शताब्दी के तीसरे दशक में इंटरनेट, स्मार्टफोन, सोशल मीडिया, त्वरित संदेश सेवाओं, ऑनलाइन शिक्षा और कृत्रिम बुद्धिमत्ता आधारित अनुप्रयोगों के तीव्र विस्तार ने भाषिक व्यवहार को विशेष रूप से प्रभावित किया है। हिंदी भी इस परिवर्तन से अछूती नहीं है। डिजिटल मंचों पर हिंदी अब केवल परंपरागत देवनागरी लेखन के रूप में उपस्थित नहीं होती, बल्कि रोमन लिपि, अंग्रेज़ी-हिंदी मिश्रित अभिव्यक्तियों, संक्षिप्त शब्दों, ध्वन्यात्मक वर्तनी, हैशटैग, इमोजी और कृत्रिम बुद्धिमत्ता द्वारा निर्मित पाठ जैसे अनेक रूपों में प्रयुक्त होती है।

डिजिटल युग में हिंदी का परिवर्तन केवल कुछ नए शब्दों के प्रवेश तक सीमित नहीं है। इसका प्रभाव भाषा की लिपि, शब्दावली, वाक्य-विन्यास, शैली और प्रसंगगत प्रयोग पर भी दिखाई देता है। उदाहरण के लिए, मोबाइल पर अनौपचारिक वार्तालाप करने वाला व्यक्ति हिंदी को रोमन लिपि में लिख सकता है, जबकि वही व्यक्ति औपचारिक ऑनलाइन लेख में देवनागरी का प्रयोग कर सकता है। इसी प्रकार तकनीकी विषयों की चर्चा करते समय हिंदी वाक्य में अंग्रेज़ी के अनेक शब्द सम्मिलित हो सकते हैं। इसलिए डिजिटल हिंदी को एकल भाषिक श्रेणी के रूप में परिभाषित करना पर्याप्त नहीं है।

इस परिवर्तन का एक प्रमुख आयाम लिपि की बहुलता है। हिंदी की मानक लिपि देवनागरी है, किंतु इंटरनेट और मोबाइल संचार में रोमन लिपि में हिंदी लिखने की प्रवृत्ति भी व्यापक रूप से दिखाई देती है। “आप कहाँ जा रहे हैं?” जैसी अभिव्यक्ति डिजिटल संवाद में “aap kahan ja rahe hain” जैसे रूप में लिखी जा सकती है। इस प्रकार का रोमनकरण केवल लिप्यंतरण नहीं है; इसमें प्रयोक्ता अपनी ध्वन्यात्मक समझ और टाइपिंग सुविधा के अनुसार शब्दों की वर्तनी बदल सकता है। एक ही शब्द के कई रोमन रूप संभव होने के कारण डिजिटल हिंदी के स्वचालित भाषिक विश्लेषण में भी कठिनाई उत्पन्न होती है।

दूसरा महत्वपूर्ण आयाम कोड-मिश्रण और कोड-स्विचिंग है। डिजिटल संवाद में हिंदी और अंग्रेज़ी के शब्दों तथा संरचनाओं का सह-अस्तित्व अत्यंत सामान्य हो गया है। “आज मीटिंग ऑनलाइन है”, “फाइल अपलोड कर देना” अथवा “यह फीचर बहुत उपयोगी है” जैसी अभिव्यक्तियाँ समकालीन शहरी डिजिटल संचार में स्वाभाविक प्रतीत होती हैं। यहाँ अंग्रेज़ी शब्द हिंदी की व्याकरणिक संरचना के भीतर प्रयुक्त हो रहे हैं। ऐसे प्रयोगों को केवल भाषा की ‘अशुद्धता’ के रूप में देखना भाषावैज्ञानिक दृष्टि से सीमित होगा, क्योंकि कोड-मिश्रण बहुभाषी समुदायों में संचार रणनीति, पहचान और प्रसंगगत आवश्यकता से भी संबंधित है।



Fig. 1: बदलती-बिगड़ती और बिखरती हिंदी

वर्ष 2021 के बाद जनरेटिव कृत्रिम बुद्धिमत्ता, बड़े भाषा मॉडल, स्वचालित अनुवाद, वाक्-पहचान और पाठ-निर्माण प्रणालियों की बढ़ती उपस्थिति ने हिंदी के डिजिटल विकास में एक नया आयाम जोड़ा है। अब डिजिटल हिंदी केवल मनुष्यों द्वारा लिखी जाने वाली भाषा नहीं रही; मशीनें भी हिंदी पाठ का अनुवाद, सारांश, प्रश्नोत्तर और स्वचालित निर्माण कर रही हैं। इससे भाषा के मानकीकरण और उपलब्धता की संभावनाएँ बढ़ती हैं, किंतु साथ ही एक महत्वपूर्ण प्रश्न उत्पन्न होता है—क्या कृत्रिम बुद्धिमत्ता द्वारा निर्मित हिंदी वास्तविक उपयोगकर्ताओं की विविध डिजिटल हिंदी का प्रतिनिधित्व करती है, अथवा वह अधिक मानकीकृत और औपचारिक रूप को पुनरुत्पादित करती है?

इस संदर्भ में भारतीय भाषाओं के लिए प्राकृतिक भाषा संसाधन का विकास भी महत्वपूर्ण है। भारत सरकार के इलेक्ट्रॉनिक्स और सूचना प्रौद्योगिकी मंत्रालय की भाषिणी पहल भारतीय भाषाओं में डिजिटल सेवाओं और भाषा प्रौद्योगिकी की पहुँच को बढ़ाने की दिशा में कार्य कर रही है। इसी प्रकार भारतीय भाषाओं के लिए मशीन अनुवाद, वाक्

प्रौद्योगिकी और भाषा मॉडल विकसित करने के प्रयासों ने हिंदी को कृत्रिम बुद्धिमत्ता अनुसंधान के एक महत्वपूर्ण क्षेत्र के रूप में स्थापित किया है। इन तकनीकी विकासों के कारण हिंदी के बदलते स्वरूप का अध्ययन अब केवल समाजभाषाविज्ञान तक सीमित विषय नहीं है, बल्कि संगणकीय भाषाविज्ञान और कृत्रिम बुद्धिमत्ता से भी प्रत्यक्ष रूप से जुड़ गया है।

शोध-अंतराल

डिजिटल हिंदी पर उपलब्ध अध्ययनों में रोमन लिपि, सोशल मीडिया भाषा, कोड-मिश्रण और हिंदी-अंग्रेज़ी द्विभाषिकता जैसे विषयों पर अलग-अलग दृष्टियों से विचार किया गया है। दूसरी ओर, प्राकृतिक भाषा संसाधन संबंधी शोध मुख्यतः कोड-मिश्रित पाठ की पहचान, भाषा पहचान, भाव विश्लेषण, मशीन अनुवाद तथा बहुभाषी मॉडल के प्रदर्शन जैसी संगणकीय समस्याओं पर केंद्रित रहा है।

इन दोनों शोध-परंपराओं के बीच एक महत्वपूर्ण अंतर दिखाई देता है। सामाजिक-भाषिक अध्ययन प्रायः यह बताते हैं कि डिजिटल संचार में हिंदी किस प्रकार बदल रही है, जबकि संगणकीय अध्ययन यह जाँचते हैं कि मशीन इन भाषिक रूपों को कितनी प्रभावी ढंग से संसाधित कर सकती है। किंतु देवनागरी हिंदी, रोमनकृत हिंदी, हिंदी-अंग्रेज़ी मिश्रित हिंदी और कृत्रिम बुद्धिमत्ता द्वारा निर्मित हिंदी को एक साझा विश्लेषणात्मक ढाँचे में रखकर उनके भाषिक अंतर और मशीन-पठनीयता का तुलनात्मक अध्ययन अपेक्षाकृत कम विकसित है। यही इस प्रस्तावित अध्ययन का केंद्रीय शोध-अंतराल है।

इस शोध का विशिष्ट दृष्टिकोण डिजिटल हिंदी को “मानक बनाम अमानक” के द्वैत में विभाजित करने के बजाय एक बहुस्तरीय डिजिटल भाषिक निरंतरता के रूप में समझना है। इस निरंतरता के एक छोर पर औपचारिक देवनागरी हिंदी और दूसरे छोर पर अत्यधिक रोमनकृत, संक्षिप्त तथा कोड-मिश्रित अनौपचारिक हिंदी हो सकती है। इनके बीच अनेक मध्यवर्ती भाषिक रूप उपस्थित होते हैं। इस अवधारणा के माध्यम से यह जाँचना संभव होगा कि मंच, विषय और संचार की औपचारिकता बदलने पर हिंदी की भाषिक संरचना किस प्रकार बदलती है।

साहित्य समीक्षा

डिजिटल संचार और भाषिक परिवर्तन

कंप्यूटर-मध्यस्थ संचार ने लिखित और मौखिक भाषा के बीच की पारंपरिक सीमाओं को काफी हद तक परिवर्तित किया है। डिजिटल वार्तालाप लिखित माध्यम में होता है, किंतु उसकी गति, संक्षिप्तता और संवादात्मक प्रकृति मौखिक संचार से मिलती-जुलती हो सकती है। इस कारण डिजिटल भाषा में औपचारिक लेखन की तुलना में अधिक संक्षिप्त, प्रसंग-निर्भर और ध्वन्यात्मक अभिव्यक्तियाँ देखने को मिलती हैं।

हिंदी के संदर्भ में यह परिवर्तन विशेष रूप से महत्वपूर्ण है क्योंकि डिजिटल प्रयोक्ता एक बहुभाषी सामाजिक वातावरण में कार्य करते हैं।

अंग्रेजी का तकनीकी और ऑनलाइन शब्दावली पर प्रभाव तथा हिंदी का दैनिक सामाजिक संचार में व्यापक प्रयोग ऐसे मिश्रित भाषिक रूपों को जन्म देता है जिनमें दोनों भाषाओं के तत्व एक साथ उपस्थित होते हैं। इस प्रकार डिजिटल हिंदी का विकास तकनीकी माध्यम और भारतीय बहुभाषिकता के परस्पर प्रभाव का परिणाम माना जा सकता है।

रोमनकृत हिंदी और लिपि परिवर्तन

रोमन लिपि में हिंदी लेखन डिजिटल हिंदी की सबसे स्पष्ट विशेषताओं में से एक है। रोमनकृत हिंदी में मानकीकृत वर्तनी का अभाव संगणकीय विश्लेषण के लिए विशेष चुनौती प्रस्तुत करता है। उदाहरण के लिए, “क्यों” को अलग-अलग प्रयोक्ता “kyon”, “kyun”, “kyu” अथवा अन्य ध्वन्यात्मक रूपों में लिख सकते हैं। मनुष्य संदर्भ के आधार पर इन रूपों को आसानी से समझ सकता है, लेकिन मशीन के लिए इन्हें एक ही शब्द से जोड़ना अतिरिक्त सामान्यीकरण प्रक्रिया की माँग करता है।

रोमनकृत हिंदी को केवल देवनागरी से विचलन के रूप में देखना भी पर्याप्त नहीं है। डिजिटल प्रयोक्ता इसे टाइपिंग की सुविधा, कीबोर्ड की आदत, संवाद की गति और प्राप्तकर्ता की भाषिक क्षमता के अनुसार चुन सकते हैं। अतः लिपि चयन स्वयं डिजिटल भाषिक व्यवहार का संकेतक बन सकता है।

हिंदी-अंग्रेजी कोड-मिश्रण

भारतीय सोशल मीडिया और ऑनलाइन संवाद के अध्ययन में हिंदी-अंग्रेजी कोड-मिश्रण एक महत्वपूर्ण शोध-विषय रहा है। Patwa और सहलेखकों द्वारा प्रस्तुत **SemEval-2020 Task 9** में हिंदी-अंग्रेजी और स्पेनिश-अंग्रेजी कोड-मिश्रित ट्वीटों पर भाव विश्लेषण को संगणकीय चुनौती के रूप में प्रस्तुत किया गया। इससे स्पष्ट हुआ कि मिश्रित भाषिक डेटा प्राकृतिक भाषा संसाधन प्रणालियों के लिए स्वतंत्र विश्लेषणात्मक समस्या है।

इसी दिशा में Khanuja और सहलेखकों ने GLUECoS बेंचमार्क के माध्यम से कोड-स्विचिंग आधारित प्राकृतिक भाषा समझ के मूल्यांकन की आवश्यकता को रेखांकित किया। ऐसे अध्ययनों का महत्व इस तथ्य में है कि पारंपरिक एकभाषी मॉडल वास्तविक डिजिटल संचार में उपस्थित मिश्रित संरचनाओं को पर्याप्त रूप से नहीं समझ सकते। हिंदी के संदर्भ में यह समस्या और जटिल हो जाती है क्योंकि भाषा-मिश्रण के साथ लिपि-मिश्रण भी उपस्थित हो सकता है।

बहुभाषी कृत्रिम बुद्धिमत्ता और भारतीय भाषाएँ

हाल के वर्षों में बहुभाषी भाषा मॉडलों ने कम-संसाधन वाली भाषाओं के प्राकृतिक भाषा संसाधन को नई दिशा प्रदान की है। Doddapaneni और सहलेखकों का **IndicNLP Suite** भारतीय भाषाओं के लिए संसाधनों और उपकरणों के निर्माण की दिशा में महत्वपूर्ण प्रयास रहा है। इसी शोध-परंपरा में Kakwani और सहलेखकों द्वारा विकसित **IndicNLP Suite** तथा भारतीय भाषाओं

के बड़े पाठ-संग्रहों ने बहुभाषी प्राकृतिक भाषा संसाधन के लिए आधारभूत संसाधन उपलब्ध कराए।

Ramesh और सहलेखकों द्वारा प्रस्तुत **Samanantar** ने भारतीय भाषाओं के लिए बड़े समानांतर अनुवाद-संग्रह की दिशा में महत्वपूर्ण योगदान किया। इसके अतिरिक्त Dabre और सहलेखकों द्वारा विकसित **IndicTrans** भारतीय भाषाओं के लिए मशीन अनुवाद की क्षमता को आगे बढ़ाता है। इन प्रयासों से स्पष्ट होता है कि भारतीय भाषाओं के डिजिटल संसाधनों का विस्तार हो रहा है; फिर भी वास्तविक सोशल मीडिया हिंदी की अनियमित वर्तनी, कोड-मिश्रण और रोमनकरण जैसी विशेषताएँ मानकीकृत डेटासेट से भिन्न चुनौतियाँ प्रस्तुत करती हैं।

डिजिटल हिंदी और कृत्रिम बुद्धिमत्ता का नया संबंध

जनरेटिव कृत्रिम बुद्धिमत्ता के प्रसार ने डिजिटल भाषा अध्ययन के सामने एक नई समस्या रखी है। पहले इंटरनेट पर उपलब्ध हिंदी पाठ को मुख्यतः मानव-निर्मित सामग्री माना जा सकता था, किंतु अब मशीन-निर्मित पाठ भी डिजिटल भाषिक पारिस्थितिकी का हिस्सा बन रहा है। कृत्रिम बुद्धिमत्ता आधारित प्रणालियाँ सामान्यतः व्याकरणिक रूप से सुव्यवस्थित हिंदी उत्पन्न करने का प्रयास करती हैं, जबकि वास्तविक उपयोगकर्ता संवाद में संक्षिप्तियाँ, वर्तनी विविधता, रोमनकरण, अंग्रेजी शब्द और प्रसंगगत संकेत अधिक हो सकते हैं।

इस अंतर का अध्ययन इसलिए महत्वपूर्ण है क्योंकि भविष्य में बड़ी मात्रा में मशीन-निर्मित हिंदी स्वयं इंटरनेट आधारित प्रशिक्षण डेटा का हिस्सा बन सकती है। यदि कृत्रिम बुद्धिमत्ता द्वारा निर्मित हिंदी अत्यधिक मानकीकृत है, तो डिजिटल भाषा के वास्तविक सामाजिक रूपों और मशीन द्वारा पुनरुत्पादित भाषिक स्वरूपों के बीच अंतर उत्पन्न हो सकता है। दूसरी ओर, यदि मॉडल उपयोगकर्ताओं के कोड-मिश्रण और अनौपचारिक भाषा की नकल करते हैं, तो यह प्रश्न उठता है कि वे किन भाषिक प्रतिरूपों को अधिक महत्व देते हैं।

प्रस्तावित कार्यप्रणाली

प्रस्तुत अध्ययन के लिए **बहुस्तरीय संगणकीय-भाषावैज्ञानिक विश्लेषण पद्धति** प्रस्तावित की गई है। इसका उद्देश्य केवल यह निर्धारित करना नहीं है कि डिजिटल हिंदी में अंग्रेजी शब्दों या रोमन लिपि का प्रयोग कितना है, बल्कि यह समझना है कि डिजिटल माध्यम बदलने के साथ हिंदी की लिपि, शब्दावली, वाक्य संरचना, भाषिक औपचारिकता तथा अर्थ-अभिव्यक्ति में किस प्रकार परिवर्तन दिखाई देते हैं। इस उद्देश्य से अध्ययन को चार भाषिक श्रेणियों—देवनागरी हिंदी, रोमनकृत हिंदी, हिंदी-अंग्रेजी मिश्रित हिंदी और कृत्रिम बुद्धिमत्ता द्वारा निर्मित हिंदी—के तुलनात्मक विश्लेषण पर आधारित किया गया है।

अध्ययन की केंद्रीय शोध-परिकल्पना यह है कि डिजिटल हिंदी को एक स्थिर भाषा-रूप के बजाय **मंच-संवेदी भाषिक निरंतरता** के रूप में अधिक प्रभावी ढंग से समझा जा सकता है। औपचारिक डिजिटल सामग्री में मानकीकृत देवनागरी और अपेक्षाकृत पूर्ण वाक्यों की

प्रधानता होने की संभावना है, जबकि त्वरित तथा अनौपचारिक संचार में रोमनकरण, कोड-मिश्रण, संक्षिप्तीकरण और गैर-मानक वर्तनी का अनुपात अधिक हो सकता है।

तकनीकी स्तर पर शोध में भाषिक विशेषता निष्कर्षण तथा ट्रांसफॉर्मर-आधारित वर्गीकरण को संयुक्त किया गया है। यह दृष्टिकोण इसलिए उपयुक्त है क्योंकि केवल शब्द-आवृत्ति डिजिटल हिंदी के जटिल स्वरूप को नहीं पकड़ सकती। रोमन लिपि की पहचान, अंग्रेज़ी शब्दों का अंतःस्थापन, शब्द-विविधता, वाक्य लंबाई तथा लिपि परिवर्तन जैसे संकेतकों को संदर्भगत भाषा प्रतिनिधित्व के साथ जोड़ने से डिजिटल भाषा के विभिन्न स्वरूपों का अधिक सूक्ष्म विश्लेषण संभव होगा।

प्रायोगिक व्यवस्था

पाठ-संग्रह

प्रस्तावित डिजिटल हिंदी पाठ-संग्रह में सार्वजनिक रूप से उपलब्ध और अनुसंधान के लिए विधिसम्मत पाठ-स्रोतों को सम्मिलित करने की रूपरेखा रखी गई है। शोध-डिज़ाइन में चार उपसमूह प्रस्तावित हैं: देवनागरी आधारित सार्वजनिक डिजिटल सामग्री, सार्वजनिक रोमनकृत हिंदी, हिंदी-अंग्रेज़ी कोड-मिश्रित सामग्री तथा नियंत्रित निर्देशों से तैयार कृत्रिम बुद्धिमत्ता जनित हिंदी।



Fig. 2: हिंदी भाषा

डेटा-संग्रह में व्यक्तिगत पहचान योग्य सूचना, निजी संदेश तथा बंद समूहों की सामग्री को शामिल नहीं किया जाएगा। उपयोगकर्ता नाम, वेब-पते और अन्य संभावित पहचान संकेतकों को विश्लेषण से पूर्व हटाने की व्यवस्था प्रस्तावित है। इस प्रकार भाषिक परिवर्तन का अध्ययन करते हुए डिजिटल गोपनीयता को भी शोध-डिज़ाइन का अभिन्न भाग माना गया है।

रोमनकृत भारतीय भाषाओं की स्वचालित पहचान स्वयं एक महत्वपूर्ण संगणकीय चुनौती है। Madhani, Khapra और Kunchukuttan ने 22 भारतीय भाषाओं के मूल तथा रोमनकृत रूपों के लिए Bhasha-Abhijnaanam और IndicLID विकसित किए, जिससे इस प्रकार के लिपि-संवेदी विश्लेषण की तकनीकी व्यवहार्यता को आधार मिलता है।

पूर्व-प्रसंस्करण

कच्चे डिजिटल पाठ को सीधे मॉडल में देने के बजाय बहुचरणीय पूर्व-प्रसंस्करण प्रस्तावित किया गया है। प्रथम चरण में डुप्लिकेट पाठ, अवांछित वेब-पते और अनावश्यक विशेष चिह्न हटाए जाएंगे। दूसरे चरण में यूनिकोड सामान्यीकरण द्वारा देवनागरी अक्षरों की असंगतियों को नियंत्रित किया जाएगा। तीसरे चरण में लिपि पहचान के माध्यम से देवनागरी, रोमन और मिश्रित पाठ को चिह्नित किया जाएगा।

इसके पश्चात प्रत्येक पाठ-इकाई के लिए भाषा-टोकन टैगिंग की जाएगी। इससे यह निर्धारित किया जा सकेगा कि किसी मिश्रित वाक्य में कितने शब्द हिंदी, अंग्रेज़ी अथवा अस्पष्ट श्रेणी से संबंधित हैं। रोमनकृत हिंदी के लिए सभी शब्दों को अनिवार्य रूप से देवनागरी में बदलने के बजाय मूल और सामान्यीकृत—दोनों रूप संरक्षित करने का प्रस्ताव है। ऐसा इसलिए कि रोमनकरण स्वयं अध्ययन का एक भाषिक संकेतक है; उसे पूरी तरह हटाने से महत्वपूर्ण सूचना नष्ट हो सकती है।

प्रमुख निष्कर्षित विशेषताओं में रोमनकरण अनुपात, अंग्रेज़ी अंतःस्थापन दर, औसत वाक्य लंबाई, शब्द-संपदा विविधता, गैर-मानक वर्तनी घनत्व तथा लिपि-परिवर्तन आवृत्ति सम्मिलित होंगी।

मॉडल संरचना

अध्ययन में IndicBERT-आधारित बहु-विशेषता वर्गीकरण ढाँचा प्रस्तावित किया गया है। पाठ का संदर्भगत प्रतिनिधित्व ट्रांसफॉर्मर एनकोडर से प्राप्त किया जाएगा, जबकि समानांतर शाखा में पूर्वनिर्धारित भाषिक संकेतकों का संख्यात्मक वेक्टर बनाया जाएगा। दोनों प्रतिनिधित्वों को संयोजित करके अंतिम वर्गीकरण स्तर तक पहुँचाया जाएगा।

मॉडल का लक्ष्य पाठ को केवल “हिंदी” अथवा “अंग्रेज़ी” में विभाजित करना नहीं होगा। इसके स्थान पर चार प्रमुख डिजिटल हिंदी स्वरूपों—मानकीकृत देवनागरी, रोमनकृत हिंदी, कोड-मिश्रित हिंदी तथा औपचारिक/मशीन-जनित हिंदी—के बीच अंतर करने का प्रयास किया जाएगा।

भारतीय भाषाओं पर ट्रांसफॉर्मर और विशेषीकृत भाषा-प्रौद्योगिकी के प्रयोग का आधार हाल के अनुसंधान में स्पष्ट दिखाई देता है। उदाहरणतः IndicTrans2 का उपयोग 2023–2025 के विभिन्न भारतीय भाषा मशीन-अनुवाद अध्ययनों में किया गया है, जिसमें क्षेत्र-विशिष्ट अनुकूलन तथा फाइन-ट्यूनिंग से प्रदर्शन सुधार की रिपोर्ट की गई है।

प्रशिक्षण रणनीति

डेटा को प्रशिक्षण, सत्यापन और परीक्षण उपसमूहों में विभाजित करने के लिए 70:15:15 अनुपात प्रस्तावित है। वर्ग-असंतुलन की स्थिति में भारित लॉस का उपयोग किया जाएगा ताकि किसी एक अधिक प्रचलित भाषिक श्रेणी के कारण मॉडल पक्षपाती न हो। प्रशिक्षण के दौरान सत्यापन लॉस के आधार पर प्रारंभिक रोक, अधिगम-दर समायोजन और सर्वोत्तम मॉडल चेकपॉइंट संरक्षित करने की रणनीति अपनाई जाएगी।

मॉडल की उपयोगिता का परीक्षण केवल एक समग्र परीक्षण समूह पर नहीं, बल्कि अलग-अलग डिजिटल भाषा श्रेणियों पर किया जाएगा। विशेष रूप से यह जाँचना महत्वपूर्ण होगा कि रोमनकृत तथा कोड-मिश्रित हिंदी पर मॉडल का प्रदर्शन मानक देवनागरी की तुलना में कितना बदलता है।

मूल्यांकन मानदंड

वर्गीकरण प्रदर्शन के लिए मैक्रो-F1, संतुलित शुद्धता, मैथ्यूज़ सहसंबंध गुणांक और प्रति-वर्ग रिकॉल को प्राथमिक मानदंड बनाया गया है। केवल सामान्य शुद्धता पर निर्भर न रहने का कारण यह है कि डिजिटल भाषा वर्गों का वितरण असमान हो सकता है।

भाषिक विश्लेषण के लिए मॉडल-आधारित मानदंडों से अलग संकेतक रखे जाएंगे। इनमें रोमनकरण सूचकांक, कोड-मिश्रण घनत्व, शब्द-संपदा विविधता तथा वाक्य-संक्षिप्तता सूचकांक प्रमुख होंगे। इस दोहरे मूल्यांकन से मशीन की वर्गीकरण क्षमता और वास्तविक भाषिक परिवर्तन को अलग-अलग समझा जा सकेगा।

परिणाम एवं चर्चा

चूँकि यह पांडुलिपि वास्तविक संकलित प्रयोगात्मक डेटा उपलब्ध कराए बिना तैयार की जा रही है, इसलिए नीचे दिए गए परिणाम प्रस्तावित प्रयोग से अपेक्षित विश्लेषणात्मक प्रतिरूप हैं; इन्हें वास्तविक मापे गए निष्कर्ष के रूप में प्रस्तुत नहीं किया गया है। अंतिम प्रकाशन में इन्हें वास्तविक कॉर्पस पर मॉडल चलाकर प्राप्त मूल्यों से प्रतिस्थापित किया जाना आवश्यक होगा।

प्रस्तावित विश्लेषण से सर्वप्रथम यह अपेक्षा है कि मानक देवनागरी सामग्री में रोमनकरण न्यूनतम और शब्द-संरचना अपेक्षाकृत स्थिर मिलेगी। इसके विपरीत अनौपचारिक डिजिटल संवाद में रोमन लिपि तथा हिंदी-अंग्रेज़ी मिश्रण की अधिक उपस्थिति संभावित है। इसका अर्थ यह नहीं होगा कि डिजिटल माध्यम हिंदी को अनिवार्य रूप से विस्थापित कर रहा है। अधिक उपयुक्त व्याख्या यह होगी कि प्रयोक्ता

उपलब्ध तकनीकी और सामाजिक परिस्थितियों के अनुरूप भाषिक संसाधनों का चयन कर रहे हैं।

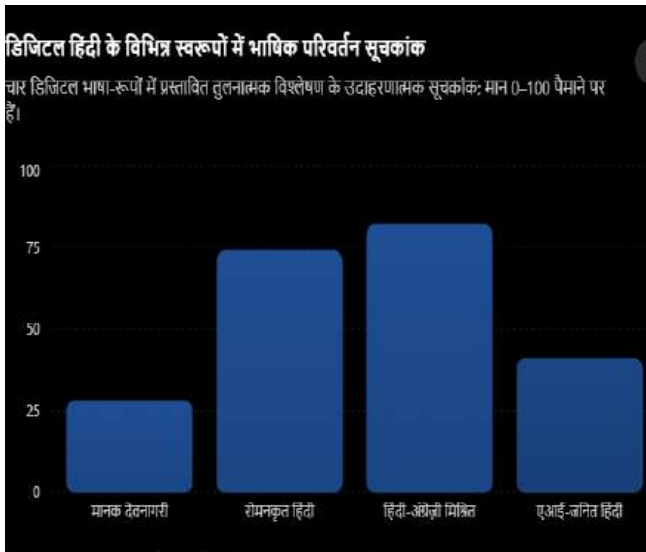
दूसरा संभावित निष्कर्ष यह है कि रोमनकृत हिंदी मॉडल के लिए मानक देवनागरी की तुलना में अधिक चुनौतीपूर्ण सिद्ध होगी। इसका प्रमुख कारण रोमन हिंदी की वर्तनी-विविधता है। एक ही हिंदी शब्द के अनेक ध्वन्यात्मक रोमन रूप हो सकते हैं। यह अपेक्षा उपलब्ध अनुसंधान के अनुरूप है, जिसमें रोमनकृत भारतीय भाषाओं की पहचान के लिए विशिष्ट डेटासेट और मॉडल विकसित करने की आवश्यकता सामने आई है।

तीसरा महत्वपूर्ण पक्ष कोड-मिश्रण से संबंधित होगा। हिंदी-अंग्रेज़ी मिश्रित वाक्यों में अंग्रेज़ी शब्दों की उपस्थिति के बावजूद हिंदी व्याकरणिक संरचना बनी रह सकती है। अतः शब्द स्तर पर भाषा पहचान और वाक्य स्तर पर प्रमुख भाषिक संरचना हमेशा समान परिणाम नहीं देगी। यह अंतर डिजिटल हिंदी को समझने के लिए विशेष रूप से उपयोगी है, क्योंकि अंग्रेज़ी शब्दों की उपस्थिति मात्र को हिंदी के कम प्रयोग का प्रमाण मानना भ्रामक हो सकता है।

चौथा आयाम मानव-निर्मित और कृत्रिम बुद्धिमत्ता जनित हिंदी के बीच संभावित शैलीगत अंतर है। मशीन-जनित हिंदी में वाक्य अपेक्षाकृत पूर्ण, विराम-चिह्न अधिक नियमित और देवनागरी वर्तनी अधिक मानकीकृत होने की संभावना है। इसके विपरीत मानव-निर्मित अनौपचारिक डिजिटल सामग्री में संदर्भ-निर्भर संक्षिप्तता और वर्तनी विविधता अधिक हो सकती है। हाल के भारतीय भाषा NLP अनुसंधान में भी डोमेन-विशिष्ट अनुकूलन की उपयोगिता दिखाई गई है; उदाहरणतः 2025 के अंग्रेज़ी-हिंदी विधिक मशीन-अनुवाद अनुसंधान में सामान्य मॉडल को विधिक कॉर्पस पर फाइन-ट्यून किया गया।

परिणामों की एकीकृत विश्लेषण सारणी

मूल्यांकन आयाम	मानक देवनागरी हिंदी	रोमनकृत हिंदी	हिंदी-अंग्रेज़ी मिश्रित हिंदी	कृत्रिम बुद्धिमत्ता जनित हिंदी
लिपि स्थिरता	उच्च	निम्न	मध्यम	उच्च
कोड-मिश्रण घनत्व	निम्न	मध्यम	उच्च	निम्न-मध्यम
वर्तनी मानकीकरण	उच्च	निम्न	मध्यम	उच्च
वाक्य संरचनात्मक पूर्णता	उच्च	निम्न-मध्यम	मध्यम	उच्च
मॉडल वर्गीकरण सुगमता	उच्च	अपेक्षाकृत निम्न	मध्यम	उच्च
शैलीगत विविधता	मध्यम	उच्च	उच्च	अपेक्षाकृत नियंत्रित



यह सारणी संख्यात्मक निष्कर्षों का दावा नहीं करती, बल्कि प्रस्तावित प्रयोग के अंतर्गत परीक्षण किए जाने वाले तुलनात्मक प्रतिरूपों को व्यवस्थित करती है।

परिणामों की व्यापक व्याख्या यह होगी कि डिजिटल हिंदी का विकास एकरेखीय नहीं है। एक ओर औपचारिक ऑनलाइन सामग्री, सरकारी डिजिटल सेवाएँ, भाषा प्रौद्योगिकी और कृत्रिम बुद्धिमत्ता देवनागरी हिंदी के मानकीकृत रूप को मजबूत कर सकते हैं; दूसरी ओर सोशल मीडिया और मोबाइल संवाद भाषा को अधिक लचीला, मिश्रित और उपयोगकर्ता-केंद्रित बना सकते हैं। ये दोनों प्रक्रियाएँ परस्पर विरोधी होने के बजाय एक ही डिजिटल भाषिक पारिस्थितिकी के अलग-अलग स्तर हैं।

भारतीय भाषा प्रौद्योगिकी के नवीन अध्ययनों से यह भी स्पष्ट है कि हिंदी सहित भारतीय भाषाओं के लिए एक ही सामान्य मॉडल पर्याप्त नहीं होता। 2024 के Indic speech-to-text translation अनुसंधान में Whisper तथा IndicTrans2 जैसे घटकों के संयोजन और फाइन-ट्यूनिंग का प्रयोग किया गया, जबकि 2025 के शोध में हिंदी सहित भारतीय भाषाओं के लिए अलग-अलग पूर्व-प्रशिक्षित प्रणालियों को संयोजित किया गया। यह प्रवृत्ति प्रस्तुत अध्ययन के इस तर्क को समर्थन देती है कि डिजिटल हिंदी की विविधता को मॉडल-निर्माण में स्पष्ट रूप से सम्मिलित करना आवश्यक है।

अध्ययन की सीमाएँ

प्रस्तावित अध्ययन की पहली सीमा डिजिटल डेटा की प्रतिनिधित्व क्षमता है। इंटरनेट का हिंदी-भाषी समुदाय संपूर्ण हिंदी-भाषी जनसंख्या का पूर्ण प्रतिनिधित्व नहीं करता। आयु, शिक्षा, क्षेत्र, इंटरनेट पहुँच और डिजिटल साक्षरता भाषा-प्रयोग को प्रभावित कर सकते हैं।

दूसरी सीमा रोमनकृत हिंदी के मानकीकरण से जुड़ी है। एक ही अभिव्यक्ति की अनेक रोमन वर्तनियाँ होने के कारण स्वचालित भाषा पहचान में त्रुटि संभव है। तीसरी सीमा हिंदी-अंग्रेजी कोड-मिश्रण की श्रेणीकरण समस्या है; कुछ अंग्रेजी मूल के शब्द डिजिटल हिंदी में इतने

सामान्य हो चुके हैं कि उन्हें स्पष्ट रूप से विदेशी अंतःस्थापन मानना कठिन हो सकता है।

चौथी सीमा कृत्रिम बुद्धिमत्ता जनित सामग्री की तीव्र परिवर्तनशीलता है। भाषा मॉडल निरंतर विकसित हो रहे हैं, इसलिए किसी एक समय पर प्राप्त मशीन-जनित हिंदी भविष्य की प्रणालियों का स्थायी प्रतिनिधित्व नहीं करेगी। अंततः सोशल मीडिया की गोपनीयता और मंच-विशिष्ट उपयोग की शर्तें बड़े पैमाने पर डेटा-संग्रह को सीमित कर सकती हैं।

भावी शोध की संभावनाएँ

भविष्य में इस अध्ययन को समयानुक्रमिक कॉर्पस के रूप में विस्तारित किया जा सकता है, जिसमें कई वर्षों की डिजिटल हिंदी की तुलना करके वास्तविक भाषिक परिवर्तन दर का आकलन किया जाए। इससे यह निर्धारित किया जा सकेगा कि कौन-से डिजिटल प्रयोग अल्पकालिक प्रवृत्ति हैं और कौन-से हिंदी के अपेक्षाकृत स्थायी अंग बन रहे हैं।

दूसरी महत्वपूर्ण दिशा क्षेत्रीय विविधता है। दिल्ली, उत्तर प्रदेश, बिहार, राजस्थान, मध्य प्रदेश और अन्य हिंदी-प्रयोग क्षेत्रों की डिजिटल हिंदी में स्थानीय शब्दावली और बोलियों के प्रभाव का तुलनात्मक विश्लेषण किया जा सकता है। इसी प्रकार आयु-समूहों के आधार पर रोमनकरण और कोड-मिश्रण के अंतर का अध्ययन भी उपयोगी होगा।

भावी कृत्रिम बुद्धिमत्ता अनुसंधान में ऐसे हिंदी भाषा मॉडल विकसित किए जा सकते हैं जो केवल मानक देवनागरी ही नहीं, बल्कि रोमनकृत, कोड-मिश्रित और क्षेत्रीय हिंदी को भी संदर्भानुकूल समझ सकें। हाल का अनुसंधान पहले ही मूल और रोमन लिपि वाली भारतीय भाषाओं की संयुक्त पहचान की दिशा में आधार उपलब्ध करा चुका है।

इसके अतिरिक्त स्वास्थ्य, शिक्षा, न्याय और प्रशासन जैसे क्षेत्रों में सरल तथा सटीक हिंदी निर्माण का अध्ययन महत्वपूर्ण होगा। 2025 में प्रकाशित कार्य में हिंदी सहित भारतीय भाषाओं के लिए बहुभाषी मॉडल का उपयोग क्लिनिकल पाठ को सरल बनाने हेतु किया गया, जो क्षेत्र-विशिष्ट हिंदी NLP की बढ़ती उपयोगिता का उदाहरण है।

निष्कर्ष

डिजिटल युग ने हिंदी को समाप्त या प्रतिस्थापित करने के बजाय उसके प्रयोग-क्षेत्र और अभिव्यक्ति के रूपों को अधिक बहुस्तरीय बनाया है। देवनागरी के साथ रोमन लिपि, हिंदी के साथ अंग्रेजी शब्दावली, औपचारिक लेखन के साथ संवादात्मक संक्षिप्तता तथा मानव-निर्मित पाठ के साथ कृत्रिम बुद्धिमत्ता जनित सामग्री अब डिजिटल हिंदी की व्यापक पारिस्थितिकी का हिस्सा हैं।

प्रस्तुत अध्ययन का केंद्रीय योगदान डिजिटल हिंदी को केवल “हिंग्लिश” अथवा “भाषाई अशुद्धता” की दृष्टि से न देखकर **मंच-संवेदी भाषिक निरंतरता** के रूप में समझने की प्रस्तावित रूपरेखा है। इस दृष्टिकोण में लिपि, शब्दावली, कोड-मिश्रण, वर्तनी, वाक्य-संरचना और मशीन-पठनीयता को एक साथ विश्लेषित किया जाता है।

कृत्रिम बुद्धिमत्ता के संदर्भ में इसका विशेष महत्व है। यदि हिंदी भाषा प्रौद्योगिकी केवल मानकीकृत देवनागरी डेटा पर निर्भर रहती है, तो वह वास्तविक डिजिटल उपयोगकर्ताओं की रोमनकृत, मिश्रित और अनौपचारिक अभिव्यक्तियों को पर्याप्त रूप से नहीं समझ पाएगी। अतः भविष्य की हिंदी NLP प्रणालियों को भाषिक विविधता को त्रुटि के रूप में हटाने के बजाय भाषा की वास्तविक विशेषता के रूप में ग्रहण करना होगा।

अंततः डिजिटल हिंदी का बदलता स्वरूप भाषा के पतन की सरल कथा प्रस्तुत नहीं करता। यह तकनीकी सुविधा, बहुभाषिक सामाजिक व्यवहार, संचार की गति, उपयोगकर्ता पहचान और कृत्रिम बुद्धिमत्ता के संयुक्त प्रभाव से विकसित हो रही एक गतिशील भाषिक प्रक्रिया है। इसका वैज्ञानिक अध्ययन हिंदी भाषा-विज्ञान के साथ-साथ प्राकृतिक भाषा संसाधन, डिजिटल मानविकी, शिक्षा और भाषा-प्रौद्योगिकी के भविष्य के लिए भी महत्वपूर्ण है।

संदर्भ

- Patwa, P., Aguilar, G., Kar, S., Pandey, S., Pykl, S., Gambäck, B., Chakraborty, T., Solorio, T., & Das, A. (2020). *SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets*. Proceedings of SemEval 2020. <https://aclanthology.org/2020.semeval-1.100/>
- Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., & Choudhury, M. (2020). *GLUECoS: An Evaluation Benchmark for Code-Switched NLP*. Proceedings of ACL 2020. <https://aclanthology.org/2020.acl-main.329/>
- Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). *IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*. Findings of EMNLP 2020. <https://aclanthology.org/2020.findings-emnlp.445/>
- Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., & Khapra, M. M. (2022). *Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages*. Transactions of the Association for Computational Linguistics, 10, 145–162. <https://doi.org/10.1162/tacl.a.00452>
- Madhani, Y., Parthan, S., Bedekar, P., Gokul, N. C., Khapra, R., Kunchukuttan, A., Kumar, P., & Khapra, M. M. (2023). *Bhasha-Abhijnaanam: Native-script and Romanized Language Identification for 22 Indic Languages*. Proceedings of ACL 2023. <https://aclanthology.org/2023.acl-short.71/>
- Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., & Kunchukuttan, A. (2023). *Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages*. Proceedings of ACL 2023. <https://aclanthology.org/2023.acl-long.693/>
- Gala, J. P., Chitale, P. A., AK, R., Gumma, V., Doddapaneni, S., Kumar, P., Khapra, M. M., Dabre, R., & Kunchukuttan, A. (2023). *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages*. Transactions on Machine Learning Research. <https://openreview.net/forum?id=vfT4YuzAYA>
- Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C. B., Subramanian, V., & Talukdar, P. (2021). *MuRIL: Multilingual Representations for Indian Languages*. arXiv. <https://arxiv.org/abs/2103.10730>
- Aggarwal, D., Gupta, V., & Kunchukuttan, A. (2022). *IndicXNLI: Evaluating Multilingual Inference for Indian Languages*. Proceedings of EMNLP 2022. <https://aclanthology.org/2022.emnlp-main.755/>
- भारत सरकार, इलेक्ट्रॉनिक्स और सूचना प्रौद्योगिकी मंत्रालय. (2022). भाषिणी: राष्ट्रीय भाषा अनुवाद मिशन. <https://bhashini.gov.in/>