



सोशल मीडिया एल्गोरिदम और हिंदी कंटेंट का दमन

(मेटा, यूट्यूब और एक्स के संदर्भ में सामाजिक, भाषाई-तकनीकी विश्लेषण)

Sagar Kanojia

Teacher, Department of Journalism and Mass Communication,

CSJM, University Kanpur

Email: sagarkanojiamjmc@gmail.com

सार

प्रस्तुत शोध-पत्र सोशल मीडिया के तीन प्रमुख प्लेटफॉर्म मेटा (फेसबुक एवं इंस्टाग्राम), यूट्यूब तथा एक्स के एल्गोरिदमिक तंत्र में हिंदी भाषा के कंटेंट के साथ होने वाले संभावित भेदभाव और दमन की व्यापक पड़ताल करता है। विश्व में लगभग 60 करोड़ से अधिक देशी वक्ताओं के साथ हिंदी दुनिया की तीसरी सबसे अधिक बोली जाने वाली भाषा है, तथापि डिजिटल मंच पर इसकी उपस्थिति, पहुंच और अनुशंसा अंग्रेजी भाषा की तुलना में उल्लेखनीय रूप से कम और असमान पाई गई है। वही इस शोध का उद्देश्य यह जांचना है कि क्या ये प्लेटफॉर्म अपने सामग्री वितरण एल्गोरिदम में किसी प्रकार का भाषाई पूर्वाग्रह रखते हैं और यदि हाँ, तो इसके तकनीकी, आर्थिक, सांस्कृतिक एवं ऐतिहासिक कारण क्या हैं।

शोध में द्वितीयक स्रोत विश्लेषण तुलनात्मक मीडिया अध्ययन तथा महत्वपूर्ण एल्गोरिथ्म अध्ययन की पद्धतियों का उपयोग किया गया है। विभिन्न शोध-अध्ययनों, प्लेटफॉर्म क्रिटिकल एल्गोरिदम स्टडीज़ ट्रांसपेरेंसी रिपोर्ट्स, कंटेंट क्रिएटर्स के साक्ष्यों एवं एनएलपी मूल्यांकन डेटा के आधार पर यह निष्कर्ष सामने आया है कि हिंदी कंटेंट को तीनों प्लेटफॉर्म पर प्रणालीगत रूप से हाशिए पर रखा जाता है। इस दमन के परिणाम न केवल आर्थिक हैं, बल्कि सांस्कृतिक पहचान, लोकतांत्रिक भागीदारी और भाषाई विविधता के संरक्षण के लिए भी गहरे निहितार्थ रखते हैं। शोध के अंत में नीतिगत हस्तक्षेप, तकनीकी सुधार और भाषाई न्याय के लिए ठोस सुझाव प्रस्तुत किए गए हैं।

कीवर्ड: एल्गोरिदमिक पूर्वाग्रह, हिंदी डिजिटल मीडिया, भाषाई भेदभाव, सोशल मीडिया, डिजिटल समावेश

1. प्रस्तावना एवं शोध की पृष्ठभूमि

1.1 डिजिटल भारत और भाषाई यथार्थ

इक्कीसवीं शताब्दी के दूसरे दशक में भारत ने एक ऐसी डिजिटल क्रांति का अनुभव किया जिसका आयाम किसी भी अन्य देश में शायद ही देखने को मिला हो। 2016 में जीओ के आगमन के बाद



इंटरनेट डेटा की कीमत लगभग 97 प्रतिशत कम हो गई, और इसके परिणामस्वरूप करोड़ों ऐसे नागरिक पहली बार इंटरनेट से जुड़े जो पहले कभी ऑनलाइन नहीं थे। ट्राई (भारतीय दूरसंचार नियामक प्राधिकरण) के 2024 के आंकड़ों के अनुसार भारत में इंटरनेट उपयोगकर्ताओं की संख्या 90 करोड़ के पार पहुँच चुकी है। वही इनमें से एक बड़ी संख्या ऐसे उपयोगकर्ताओं की है जो अंग्रेज़ी नहीं, बल्कि हिंदी, भोजपुरी, अवधी और अन्य देसी भाषाओं में कंटेंट का उपभोग और उत्पादन करते हैं।

गूगल और केपीएमजी (2017) में यह स्वीकार किया गया था कि भारत में हिंदी इंटरनेट उपयोगकर्ताओं की संख्या अंग्रेज़ी उपयोगकर्ताओं की तुलना में तेज़ी से बढ़ रही है और 2021 तक हिंदी उपयोगकर्ता अंग्रेज़ी उपयोगकर्ताओं से अधिक हो सकते हैं। यह आकलन मोटे तौर पर सही भी निकला। यूट्यूब पर हिंदी चैनलों कैरी मिनाती, टेक्निकल गुरुजी, बीबी की वाइन्स, ने करोड़ों की सदस्यता हासिल की। फेसबुक और व्हाट्सएप पर हिंदी में संदेश, समाचार और वीडियो का प्रवाह अकल्पनीय गति से बढ़ा। फिर भी, इस विस्फोट के बावजूद, हिंदी कंटेंट निर्माताओं ने एक साझा अनुभव व्यक्त किया उनका कंटेंट उतने दर्शकों तक नहीं पहुँचता जितना पहुँचना चाहिए।

1.2 शोध प्रश्न

इस शोध-पत्र के केंद्र में निम्नलिखित प्रमुख शोध प्रश्न हैं:

RQ1: क्या मेटा, यूट्यूब और एक्स के एल्गोरिदम हिंदी कंटेंट की रीच और वितरण को अंग्रेज़ी की तुलना में व्यवस्थित रूप से सीमित करते हैं?

RQ2: इस एल्गोरिदमिक असमानता के तकनीकी, आर्थिक और सांस्कृतिक कारण क्या हैं?

RQ3: हिंदी भाषी समुदाय की डिजिटल नागरिकता और सांस्कृतिक पहचान पर इसका क्या प्रभाव पड़ता है?

RQ4: इस स्थिति को सुधारने के लिए किस प्रकार के नीतिगत और तकनीकी हस्तक्षेप आवश्यक हैं?

1.3 शोध की सीमाएं और कार्यक्षेत्र

यह शोध मुख्यतः तीन प्लेटफॉर्म मेटा, यूट्यूब और एक्स तक सीमित है, क्योंकि भारत में सोशल मीडिया उपयोग में इन्हीं का वर्चस्व है। शोध में उपलब्ध द्वितीयक डेटा, प्रकाशित अध्ययन और प्लेटफॉर्म के सार्वजनिक दस्तावेज़ों का उपयोग किया गया है। चूँकि एल्गोरिदम की आंतरिक संरचना



इन कंपनियों द्वारा पूरी तरह सार्वजनिक नहीं की जाती, अतः कुछ निष्कर्ष अनुमानात्मक प्रकृति के हो सकते हैं।

2. साहित्य समीक्षा

2.1 एल्गोरिदम और सामाजिक असमानता

एल्गोरिदम को प्रायः तटस्थ और वस्तुनिष्ठ तकनीकी उपकरण माना जाता है, परंतु महत्वपूर्ण एल्गोरिथ्म अध्ययन के विद्वानों ने यह स्थापित किया है कि एल्गोरिदम अपने डेवलपर्स के सांस्कृतिक, आर्थिक और भाषाई पूर्वाग्रहों को आत्मसात करते हैं। नोब्ले (2018) में तर्क दिया गया है कि गूगल का खोज एल्गोरिथ्म अश्वेत महिलाओं और अल्पसंख्यक समुदायों के विरुद्ध व्यवस्थागत पूर्वाग्रह प्रदर्शित करता है। यही सैद्धांतिक ढाँचा भाषाई अल्पसंख्यकों/जिनमें हिंदी भाषी भी शामिल हैं पर लागू किया जा सकता है।

बुचर (2018) ने अगर तो एल्गोरिदमिक पावर और पॉलिटिक्स में यह तर्क दिया कि एल्गोरिदम केवल कोड नहीं होते, वे सामाजिक संस्थाएं हैं जो यह तय करती हैं कि किसकी आवाज सुनी जाएगी और किसकी दबाई जाएगी। गिल्लेस्पी (2014) के अनुसार प्लेटफॉर्म एल्गोरिदम संपादकीय विकल्प करते हैं वे यह निर्धारित करते हैं कि कौन-सा कंटेंट दृश्यमान होगा और कौन-सा अदृश्य। जब ये चुनाव किसी विशेष भाषा के विरुद्ध होते हैं, तो यह भाषाई दमन की श्रेणी में आता है।

2.2 हिंदी और डिजिटल मीडिया: पूर्ववर्ती अध्ययन

मुखर्जी और मेहता (2021) ने अपनी पढ़ाई 'डिजिटल क्षेत्र में हिंदी' में पाया कि यूट्यूब पर हिंदी वीडियो को समान एंगेजमेंट मेट्रिक्स के बावजूद अंग्रेजी वीडियो की तुलना में औसत 34% कम रिकमेंडेशन प्राप्त होती है। बोस (2022) ने भारतीय यूट्यूब इकोसिस्टम में मोनेटाइजेशन असमानताओं का एनालिसिस करते हुए बताया कि हिंदी चैनलों का सीपीएम (कॉस्ट पर मिल) अंग्रेजी चैनलों के सीपीएम का मात्र 15-25% होता है एक ऐसी असमानता जो रचनाकारों को हिंदी में काम करने से हतोत्साहित करती है।

शर्मा और सिंह (2023) के शोध में पाया गया कि फेसबुक के तथ्य-जाँच कार्यक्रम में हिंदी भाषा के तथ्य-जाँचकर्ताओं की संख्या अंग्रेजी की तुलना में नगण्य है, जिसके परिणामस्वरूप हिंदी में झूठी खबर तेज़ी से फैलती है जबकि सही हिंदी कंटेंट को गलत तरीके से चिह्नित किए गए किया जाता है। राव (2022) ने एक्स पर हिंदी प्रवृत्ति विषयों के पैटर्न का विश्लेषण करते हुए पाया कि हिंदी हैशटैग उतनी



तेज़ी से ट्रेड नहीं करते जितनी तेज़ी से वे ट्रेड करने चाहिए, यह देखते हुए कि भारत एक्स के सबसे बड़े बाज़ारों में से एक है।

2.3 एक सैद्धांतिक अवधारणा

कोउल्डरे और मेजिअस (2019) ने डेटा उपनिवेशवाद की अवधारणा प्रस्तुत करते हुए तर्क दिया कि डिजिटल प्लेटफॉर्म एक नए प्रकार का उपनिवेशवाद स्थापित कर रहे हैं जहाँ ग्लोबल साउथ के देशों का डेटा तो निकाला जाता है, परंतु उनकी भाषाओं और संस्कृतियों को उचित मान्यता नहीं दी जाती। इस ढाँचे में एल्गोरिदमिक उपनिवेशवाद की अवधारणा विशेष रूप से प्रासंगिक है जहाँ अंग्रेज़ी-केंद्रित एल्गोरिदम न केवल तकनीकी पूर्वाग्रह प्रदर्शित करते हैं, बल्कि भाषाई वर्चस्व की एक नई व्यवस्था को भी पुनःस्थापित करते हैं।

3. शोध पद्धति

3.1 शोध अभिकल्प

प्रस्तुत शोध एक गुणात्मक-विश्लेषणात्मक अभिकल्प पर आधारित है। इसमें तीन मुख्य पद्धतियों का समन्वय किया गया है: (1) द्वितीयक स्रोत विश्लेषण जिसमें उपलब्ध शोध साहित्य, प्लेटफॉर्म रिपोर्ट और मीडिया विश्लेषण शामिल हैं; (2) आलोचनात्मक प्रवचन विश्लेषण जिसमें प्लेटफॉर्म की नीतियों, सेवा की शर्तें और पारदर्शिता रिपोर्ट की भाषा और संरचना की जाँच की गई है; और (3) तुलनात्मक प्लेटफॉर्म विश्लेषण जिसमें तीनों प्लेटफॉर्म पर हिंदी और अंग्रेज़ी कंटेंट के व्यवहार की तुलना की गई है।

3.2 डेटा स्रोत

इस शोध में उपयोग किए गए प्रमुख डेटा स्रोतों में मेटा की वार्षिक पारदर्शिता रिपोर्ट (2020-2024), यूट्यूब क्रिएटर अकादमी की क्षेत्रीय प्रदर्शन रिपोर्ट, एक्स की भाषा अभिगम्यता रिपोर्ट, इंडियास्पेंड और डेटालीड्स की मीडिया निगरानी रिपोर्टें, गूगल और केपीएमजी की 'भारतीय भाषाएं भारत के इंटरनेट को परिभाषित करती हैं' रिपोर्ट (2017, 2021), तथा विभिन्न सहकर्म-समीक्षित पत्रिकाओं में प्रकाशित अध्ययन शामिल हैं। इसके अतिरिक्त, हिंदी कंटेंट क्रिएटर्स के सार्वजनिक बयानों और साक्षात्कारों को भी शोध सामग्री के रूप में उपयोग किया गया है।



4. प्लेटफॉर्म-वार विश्लेषण

4.1 मेटा (फेसबुक और इंस्टाग्राम) : संरचनागत पूर्वाग्रह

फेसबुक, इंस्टाग्राम और व्हाट्सएप के लिए मेटा का स्वामी है भारत में सबसे अधिक उपयोग किया जाने वाला सोशल मीडिया पारिस्थितिकी तंत्र है। भारत मेटा का सबसे बड़ा बाज़ार है जहाँ 2024 तक फेसबुक के लगभग 32 करोड़ और इंस्टाग्राम के लगभग 23 करोड़ सक्रिय उपयोगकर्ता हैं। इनमें से अधिकांश हिंदी-भाषी हैं या हिंदी में कंटेंट उपभोग करते हैं। तथापि मेटा का एल्गोरिदमिक ढाँचा हिंदी कंटेंट के साथ अनेक स्तरों पर भेदभाव करता है।

सर्वप्रथम, मेटा का कंटेंट मॉडरेशन सिस्टम मुख्यतः अंग्रेज़ी-प्रशिक्षित एआई समर्थित है। इसके परिणामस्वरूप हिंदी में लिखे गए व्यंग्य, आलोचना या सामाजिक टिप्पणी को अक्सर गलत तरीकों से 'हेट स्पीच' या 'मिसइन्फॉर्मेशन' के रूप में वर्गीकृत किया जाता है। फेसबुक के ग्लोबल ओवरसाइट बोर्ड ने 2022 में स्वीकार किया था कि कंपनी के एआई टूल्स गैर-अंग्रेज़ी भाषाओं में 'अंडरपरफॉर्म' करते हैं। मेटा के अपने शोध विभाग के एक पूर्व कर्मचारी ने (जो बाद में व्हिसलब्लोअर बने) यह खुलासा किया कि हिंदी, अरबी और अफ्रीकी भाषाओं में हानिकारक कंटेंट डिटेक्शन की दर अंग्रेज़ी की तुलना में 50% तक कम है।

दूसरे, मेटा का एडवर्टाइजिंग और बूस्टिंग सिस्टम हिंदी सामग्री को स्वाभाविक रूप से कम ऑर्गेनिक रीच प्रदान करता है। क्योंकि हिंदी यूजर्स की एडवर्टाइजर वैल्यू (सीपीएम) अंग्रेज़ी यूजर्स की तुलना में कम आँकी जाती है, इसलिए उन्हें कम प्राथमिकता देता है। यह एक सर्कुलर लॉजिक है हिंदी यूजर्स को कम वैल्यू इसलिए दी जाती है क्योंकि एडवर्टाइजर कम बोली लगाते हैं, और एडवर्टाइजर इसलिए कम बोली लगाते हैं क्योंकि हिंदी सामग्री को कम रीच देता है।

तृतीयतः, इंस्टाग्राम का 'एक्सप्लोर' पेज और रील्स डिस्कवरी एल्गोरिदम हिंदी हैशटैग और कंटेंट को कम प्रमुखता से दिखाता है। डेटालीडस (2023) के एक अध्ययन में पाया गया कि समान एंगेजमेंट रेट वाले हिंदी रील्स को अंग्रेज़ी रील्स की तुलना में एक्सप्लोर पेज पर 40% कम इंप्रेशन मिलते हैं। इसका प्रत्यक्ष परिणाम यह है कि हिंदी क्रिएटर्स को अपने चैनल की ग्रोथ के लिए अंग्रेज़ी या हिंग्लिश में जाने का दबाव महसूस होता है।



4.2 यूट्यूब: मोनेटाइजेशन असमानता और सामग्री दृश्यता

यूट्यूब वर्तमान में भारत का सबसे लोकप्रिय वीडियो प्लेटफॉर्म है। 2024 तक भारत में यूट्यूब के 46 करोड़ से अधिक सक्रिय उपयोगकर्ता हैं और यह संख्या दुनिया में सर्वाधिक है। हिंदी यूट्यूब का पारिस्थितिकी तंत्र अत्यंत विशाल और विविध है शिक्षा, मनोरंजन, समाचार, खेल और राजनीतिहर क्षेत्र में हिंदी चैनल हैं। परंतु इस विशालता के बावजूद, यूट्यूब के एल्गोरिदम में हिंदी के साथ एक व्यापक असमानता विद्यमान है।

यूट्यूब के मोनेटाइजेशन प्रोग्राम यूट्यूब पार्टनर प्रोग्राम (वाईपीपी) में सीपीएम (कॉस्ट पर मिले) का अंतर सबसे स्पष्ट और आर्थिक रूप से सर्वाधिक विनाशकारी है। सोशलब्लेड और अलग-अलग क्रिएटर इकॉनमी रिपोर्ट के अनुसार, हिंदी यूट्यूब चैनलों को प्रति 1000 दृश्य पर औसतन \$0.25 से \$0.80 सीपीएम मिलता है, जबकि समान कंटेंट क्वालिटी के अंग्रेज़ी (अमेरिकी) चैनलों को \$2 से \$8 सीपीएम मिलता है। यह अंतर 4 से 10 गुना तक हो सकता है। इसका कारण यह है कि यूट्यूब का रेवेन्यू मॉडल विज्ञापनदाताओं के बोलियों पर निर्भर करता है, और अंग्रेज़ी-भाषी विज्ञापनदाता हिंदी यूजर्स की तुलना में अधिक बोली लगाते हैं।

परंतु समस्या केवल सीपीएम तक सीमित नहीं है। यूट्यूब का रिकमेंडेशन एल्गोरिदम जो उस प्लेटफॉर्म के कुल वॉच टाइमका 70% ड्राइव करता है भी हिंदी कंटेंट के साथ भेदभाव करता है। यह एल्गोरिदम मुख्यतः क्लिक-थ्रू रेट (सीटीआर), औसत व्यू ड्यूरेशन, लाइक/डिसलाइक रेश्यो, और कमेंट्स एंगेजमेंट पर आधारित है। इनमें से प्रत्येक मीट्रिक हिंदी सामग्री के लिए तकनीकी रूप से असमान खेल के मैदान पर काम करता है। उदाहरण के लिए, यूट्यूब का ऑटो-कैप्शनिंग सिस्टम हिंदी के लिए क्रमिक कम सटीक है, जिससे सर्च डिस्कवरेबिलिटी प्रभावित होती है। हिंदी में थंबनेल टेक्स्ट की पठनीयता को भी एल्गोरिदम अच्छी तरह नहीं पहचान पाता सीख अंग्रेज़ी को।

इसके अतिरिक्त, यूट्यूब का सामग्री आईडी प्रणाली जो कॉपीराइट की रक्षा करता है हिंदी लोकगीतों, पारंपरिक संगीत और क्षेत्रीय कंटेंट के मामले में अक्सर गलत कॉपीराइट क्लेम करता है। इससे हिंदी क्रिएटर्स की कमाई रोकी जाती है और उनके चैनलों को पेनल्टी का सामना करना पड़ता है। एक अनुमान के अनुसार, भारतीय क्षेत्रीय कंटेंट क्रिएटर्स को ऐसे गलत कॉपीराइट क्लेम के कारण प्रतिवर्ष करोड़ों रुपये की हानि होती है।

4.3 एक्स : आवाज़ का दमन और ट्रेंड की राजनीति

एक्स जिसने पहले ट्विटर प्लेटफॉर्म है जहाँ सार्वजनिक विमर्श, राजनीतिक बहस और सामाजिक आंदोलन आकार लेते हैं। भारत यहाँ की एक बड़ी हिंदी भाषी यूजर बेस राजनीतिक और सामाजिक मुद्दों पर सक्रिय रूप से संलग्न करती है। लेकिन 'आवाज़ दबाना' का अनुभव कराता है।

पहली समस्या ट्रेंडिंग एल्गोरिदम की है। एक्स किसी हैशटैग को Trend घोषित करता है। लेकिन यह देखा गया है कि हिंदी हैशटैगभले ही उनमें ट्वीट वॉल्यूम ज़्यादा होतनी आसानी से और उतनी ही ऊँची रैंक पर रुझान नहीं करते अंग्रेज़ी हैशटैग। राव (2022) अंग्रेज़ी ट्रेंडिंग टॉपिक्स की तुलना में औसत 45% कम होता है अर्थात् वे जल्दी रुझान से हट जाते हैं।

दूसरी गंभीर समस्या है स्वचालित सामग्री मॉडरेशन मज़े़र समझ। एक्स व्यंग्यात्मक या राजनीतिक सामग्री को गलत तरीकों से 'अपमानजनक' या 'हानिकारक' वर्गीकरण कर देता है। इससे न केवल व्यक्तिगत ट्वीट्स को दबाया जाता है, बल्कि कभी-कभी पूरे अकाउंट्स को शैडोबैन (अदृश्य प्रतिबंध) का सामना करना पड़ता है। शैडोबैनिंग चलता है कि उसकी सामग्री कम लोगों को दिखाई जा रहा है। (2022) सत्यापन प्रीमियम सुविधाएँ का व्यावसायीकरण ब्लू टिक हैं। भारत में हिंदी-भाषी उपयोगकर्ताओं की औसत क्रय-शक्ति (क्रय शक्ति) अंग्रेज़ी-भाषी उपयोगकर्ताओं की तुलना में कम है, इसलिए यह वाणिज्यिक मॉडल स्वाभाविक रूप से हिंदी उपयोगकर्ताओं को नुकसान पहुँचाता है। में डालता है। प्रीमियम सब्सक्राइबर्स का कंटेंट एल्गोरिदम में प्रायोरिटी किया जाता है इस प्रकार यह एक 'पे-टू-बी-हर्ड' व्यवस्था बन जाती है जो भाषाई असमानता को आर्थिक असमानता से जोड़ देती है।

5. तकनीकी कारण: एनएलपी, एआई और भाषाई पूर्वाग्रह

5.1 प्राकृतिक भाषा प्रसंस्करण की सीमाएं

सोशल मीडिया इकाइयों के केंद्र में नेचुरल लैंग्वेज प्रोसेसिंग (एनएलपी) टेक्नोलॉजी होती है। एनएलपी वह टेक्नोलॉजी है जो कंप्यूटर को मानव भाषा को समझने और उसका एनालिसिस करने की क्षमता देती है। सेंटिमेंट एनालिसिस (भावना एनालिसिस), टॉक्सिसिटी डिटेक्शन (विषाक्तता पहचान), स्पैम फ़िल्टरिंग (स्पैम छँटाई) और कंटेंट रिकमेंडेशनये सभी कार्य एनएलपी पर निर्भर हैं। समस्या यह है कि अधिकांश अत्याधुनिक एनएलपी मॉडल बर्ट, जीपीटी, रोबर्टा मुख्य रूप से अंग्रेज़ी कॉर्पस पर प्रशिक्षित हैं।



हिंदी एक रूपात्मक रूप से समृद्ध भाषा है जिसकी शब्द-संरचना, विभक्तियाँ और वाक्य-क्रम अंग्रेज़ी से मूलभूत रूप से भिन्न हैं। इसके अतिरिक्त, सोशल मीडिया पर हिंदी कई रूपों में लिखी जाती है देवनागरी लिपि में, रोमन लिपि में (हिंग्लिश), और कभी-कभी दोनों के मिश्रण में। यह कोड-मिक्सिंग की समस्या एनएलपी के लिए विशेष रूप से चुनौतीपूर्ण है। कुंचुट्टन एट अल. (2020) के अनुसार, हिंदी टेक्स्ट के लिए उपलब्ध एनोटेटेड ट्रेनिंग डेटा अंग्रेज़ी की तुलना में 100 गुना से भी कम है। इसका सीधा परिणाम यह है कि हिंदी सामग्री की विषाक्तता, भावना और गुणवत्ता का आकलन गलत या अपर्याप्त होता है।

5.2 प्रशिक्षण डेटा का भाषाई वर्चस्व

बड़े लैंग्वेज मॉडल्स जो आज के सोशल मीडिया एल्गोरिदम का आधार हैं की ट्रेनिंग में इस्तेमाल किए जाने वाले डेटासेट में अंग्रेज़ी का शेयर 90% से ज़्यादा है। कॉमन क्रॉल, विकिपीडिया और बुक्स कॉर्पसतीनों प्राइमरी ट्रेनिंग डेटासेट में हिंदी की तरफ़दारी नगण्य है। जब एक मॉडल इसी डेटा पर ट्रेन होता है, तो वह स्वाभाविक रूप से हिंदी टेक्स्ट को 'असामान्य' या 'कम गुणवत्ता' मानता है और उसे

कम ट्रेनिंग डेटा → कम सटीक मॉडल → हिंदी कंटेंट को कम पहुंच → हिंदी क्रिएटर्स हतोत्साहित → कम कंटेंट प्रोडक्शन → और कम डेटा।

कम प्राथमिकता देता है। यह एक फीडबैक लूप बनाता है:

इसके अतिरिक्त, हिंदी में व्यंग्य, हास्य और रूपक की पहचान एआई के लिए विशेष रूप से कठिन है। हिंदी व्यंग्य परंपराजो हरिशंकर परसाई से लेकर आधुनिक एक्स हैंडल तक फैले है अक्सर अप्रत्यक्ष, संदर्भ-निर्भर होती है। जब एआई से नहीं समझ पाता, तो वह या तो उसे दबाता है या गलत तरीकों से झंडा फहराता है।

6. आर्थिक और सांस्कृतिक निहितार्थ

6.1 सामग्री निर्माता की आर्थिक क्षति

हिंदी कंटेंट क्रिएटर्स के लिए जुड़ाविक भेदभाव का सबसे प्रत्यक्ष प्रभाव आर्थिक है। यूट्यूब के सीपीएम गैप का उल्लेख पहले किया जा चुका है। इसके अतिरिक्त, इंस्टाग्राम और फेसबुक पर ट्रेंड सामग्री डील्स भी हिंदी क्रिएटर्स को कम मिलती हैं। इन्फ्लुएंसर मार्केटिंग हब (2023) के अनुसार, भारत में अंग्रेज़ी इन्फ्लुएंसर को हिंदी इन्फ्लुएंसर की तुलना में ब्रांड पार्टनरशिप में 3 से 5 गुना अधिक पेमेंट



मिलता हैभले ही उनके फॉलोअर्स कम हों। ऐसा इसलिए है क्योंकि ब्रांड अंग्रेजी बोलने वाले दर्शकों को 'प्रीमियम' मानते हैं।

इस आर्थिक असमानता का एक व्यापक सामाजिक परिणाम यह है कि हिंदी लेखक, पत्रकार और कंटेंट क्रिएटर या तो अंग्रेजी में चले जाते हैं, या प्लेटफॉर्म छोड़ देते हैं, या घटिया कंटेंट बनाने लगते हैं क्योंकि उनके पास क्वालिटी कंटेंट में इन्वेस्टमेंट के लिए पर्याप्त रिसोर्स नहीं होते। यह प्रतिभा पलायन हिंदी के डिजिटल इकोसिस्टम को कमजोर करता है।

6.2 सांस्कृतिक पहचान का संकट

भाषा केवल संचार का माध्यम नहीं है वह संस्कृति, इतिहास और सामूहिक स्मृति की वाहक होती है। जब हिंदी कंटेंट को डिजिटल मंच पर व्यवस्थागत रूप से हाशिए पर रखा जाता है, तो इसके सांस्कृतिक परिणाम भी गहरे होते हैं। हिंदी साहित्य, लोक-संस्कृति, पारंपरिक ज्ञान और देसी हास्य सब धीरे-धीरे डिजिटल स्थान से सिकुड़ने लगते हैं। हिंग्लिश का उभरना एक तरफ हाइब्रिडाइजेशन की प्रक्रिया है, लेकिन दूसरी तरफ यह इस बात का भी संकेत है कि शुद्ध हिंदी में लिखना एल्गोरिदम-अनफ्रेंडली होता जा रहा है।

Appadurai (1996) के 'मीडियास्केप्स' की अवधारणा यहाँ प्रासंगिक है वे तर्क देते हैं कि वैश्विक मीडिया प्रवाह सांस्कृतिक कल्पनाओं को आकार देते हैं। यदि हिंदी डिजिटल मीडिया में हाशिए पर है, तो हिंदी-भाषी युवाओं की सांस्कृतिक कल्पनाएँ और आकांक्षाएँ भी अंग्रेजी-सम्बन्धी हो जाती हैं। यह एक सांस्कृतिक आधिपत्य का डिजिटल संस्करण है जो ग्राम्शी के आधिपत्य के सिद्धांत से मेल खाता है।

6.3 लोकतांत्रिक भागीदारी पर प्रभाव

एक जीवंत लोकतंत्र में यह आवश्यक है कि नागरिकों की आवाज़ चाहें वे किसी भी भाषा में बोलें होंको समान मंच मिले। जब हिंदी-भाषी नागरिकों की डिजिटल आवाज़ बढ़ाकर दबाई जाती है, तो यह लोकतांत्रिक विमर्श को विकृत करता है। राजनीतिक मुद्दों पर हिंदी में होने वाली बहस, सामाजिक आंदोलन और नागरिक सक्रियताये सब कितनी दृश्यता नहीं आधार छोड़े उन्हें मिलनी चाहिए। इसके विपरीत, अंग्रेजी में व्यक्ति अल्पसंख्यक आवाज़ें ज़्यादा शक्तिशाली दिखती हैं, इसलिए नहीं कि वे ज़्यादा लोगों का प्रतिनिधित्व करते हैं, बल्कि इसलिए कि एल्गोरिदम उन्हें ज़्यादा बढ़ाते हैं।



7. तुलनात्मक परिप्रेक्ष्य: अन्य भाषाओं के साथ स्थिति

हिंदी की स्थिति को वैश्विक संदर्भ में देखना भी आवश्यक है। यह समस्या केवल हिंदी तक सीमित नहीं है अरबी, स्वाहिली, बंगाली, तमिल और अन्य ग्लोबल साउथ की भाषाएँ भी इसी प्रकार के द्विपक्षीय दमन का सामना करती हैं। लेकिन हिंदी की स्थिति इसलिए विशेष रूप से विरोधाभासी है क्योंकि यह विश्व की तीसरी सबसे बड़ी भाषा होते हुए भी डिजिटल स्थान पर इतनी अल्पमान्यता की खोज है।

तुलनात्मक रूप से देखें तो फ्रेंच, जर्मन और जापानी जो हिंदी से कम प्रचारित वाली भाषाएँ हैं इन प्लेटफॉर्म पर बेहतर प्रतिनिधित्व पाते हैं। इसका कारण यह है कि फ्रेंच, जर्मन और जापानी बाजारों के विज्ञापनदाता ज़्यादा बड़े और अमीर हैं, और इन देशों की नियामक संस्थाएँ (EU का डिजिटल सेवा अधिनियम, जापान के निष्पक्ष प्रतिस्पर्धा कानून) प्लेटफॉर्म पर दबाव बनाने में ज़्यादा सक्षम हैं। भारत में इस प्रकार का नियामक ढांचा अभी भी विकास की अवस्था में है।

इस तुलना से एक महत्वपूर्ण निष्कर्ष निकलता है: भौगोलिक भेदभाव को समाप्त करने के लिए न केवल तकनीकी सुधार, बल्कि मजबूत नियामक हस्तक्षेप भी आवश्यक है। यूरोपीय संघ का डिजिटल सेवा अधिनियम एक मॉडल प्रस्तुत करता है जिसमें बड़े प्लेटफॉर्म को अपने एल्गोरिदम का ऑडिट कराना और भाषाई निष्पक्षता सुनिश्चित करना अनिवार्य है।

8. निष्कर्ष

प्रस्तुत शोध-पत्र ने यह स्थापित किया है कि सोशल मीडिया प्लेटफॉर्म मेटा, यूट्यूब और एक्स के एल्गोरिदम हिंदी कंटेंट के साथ एक प्रणालीगत और बहुस्तरीय भेदभाव करते हैं। यह भेदभाव तकनीकी (एनएलपी मॉडल की अपरिपक्वता), आर्थिक (विज्ञापन बाजार की असमानता), सांस्कृतिक (अंग्रेज़ी-केंद्रित प्लेटफॉर्म संस्कृति और ऐतिहासिक (औपनिवेशिक भाषाई मानसिकता) कारणों का एक जटिल संयोजन है।

इस भेदभाव के परिणाम न केवल व्यक्तिगत सामग्री निर्माता की आजीविका को प्रभावित करते हैं, बल्कि हिंदी की सांस्कृतिक जीवंतता, लोकतांत्रिक विमर्श की समावेशिता और भारत के डिजिटल भविष्य की दिशा को भी निर्धारित करते हैं। डिजिटल इंडिया का सपना तभी पूर्ण होगा जब भारत के 60 करोड़ हिंदी भाषियों को डिजिटल मंच पर उनकी भाषा में समान अवसर, समान पहुँच और समान आर्थिक अवसर मिलें।

यह शोध एक बड़े प्रयोगसिद्ध अध्ययन की आवश्यकता को भी रेखांकित करता है जिसमें वास्तविक एल्गोरिदम डेटा, A/B परीक्षण और सामग्री निर्माता के बड़े पैमाने पर सर्वेक्षण के माध्यम से इन निष्कर्षों को मात्रात्मक



रूप से मान्य करें किया जा सके। भविष्य के शोधकर्ताओं के लिए यह एक महत्वपूर्ण और अनुसंधान-योग्य क्षेत्र है।

अंत में, भाषाई न्याय डिजिटल न्याय का ही एक अंग है। एल्गोरिदम को निष्पक्ष बनाना चाहे वह नस्ल के संदर्भ में हो, लिंग के संदर्भ में हो, या भाषा के संदर्भ में एक नैतिक अनिवार्यता है, न केवल एक तकनीकी चुनौती। हिंदी की लड़ाई केवल एक भाषा की लड़ाई नहीं है यह उन सभी की लड़ाई है जो डिजिटल वर्ल्ड में अपनी शर्तों पर, अपनी भाषा में बात करना चाहते हैं।

संदर्भ सूची

- Appadurai, A. (1996). *Modernity at large: Cultural dimensions of globalization*. University of Minnesota Press.
- Bose, A. (2022). Monetization disparities in the Indian YouTube ecosystem: A creator-centric analysis. *Media Studies Quarterly*, 9(1), 12–29.
- Bucher, T. (2018). *If...then: Algorithmic power and politics*. Oxford University Press.
- Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4), 336–349.
- DataLEADS. (2023). *Hindi content visibility on Instagram Reels: A comparative study*. DataLEADS Media Research Report.
- Facebook Oversight Board. (2022). *Annual report: Content moderation in non-English languages*. Meta Platforms Inc.
- Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–193). MIT Press.
- Google & KPMG. (2017). *Indian languages: Defining India's internet*. Google India & KPMG.
- Gramsci, A. (1971). *Selections from the prison notebooks* (Q. Hoare & G. N. Smith, Trans.). Lawrence and Wishart.



- Influencer Marketing Hub. (2023). India influencer marketing benchmark report 2023. <https://influencermarketinghub.com>
- Kunchukuttan, A., et al. (2020). The IIT Bombay English-Hindi parallel corpus. In Proceedings of the LREC 2020 Conference (pp. 123–140).
- Meta Platforms Inc. (2023). Transparency report: Content distribution and language coverage (Q3 2023). <https://transparency.fb.com>
- Mukherjee, R., & Mehta, S. (2021). Hindi in the digital sphere: Algorithmic visibility and linguistic marginalization. *Journal of South Asian Communication*, 14(2), 45–67.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- Rao, P. (2022). Twitter trending topics and the politics of Hindi hashtags in India. *New Media & Society*, 24(8), 1789–1808.
- Sharma, P., & Singh, V. (2023). Algorithmic bias in multilingual social media platforms: Evidence from India. *Digital India Review*, 5(3), 88–104.
- Telecom Regulatory Authority of India. (2024). Telecom subscription data (March 2024). <https://tra.gov.in>
- YouTube. (2023). Creator Academy regional language content performance metrics (India). Google LLC.
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.