पाली: फ़ारसी-अरबी लिपियों के लिए भाषाई पहचान बेंचमार्क
(PALI: A Language Identification Benchmark for Perso-Arabic Scripts)

सीना अहमदी Sina Ahmadi
मिलिंद अगर्वाल Milind Agarwal
अंतोनिओस अनस्तासोपोलोस Antonios Anastasopoulos
कंप्यूटर विज्ञान विभाग (Department of Computer Science)
जॉर्ज मेसन विश्वविद्यालय (George Mason University)
{sahmad46,magarwa,antonis}@gmu.edu

सारांश (Abstract)

फ़ारसी-अरबी लिपियाँ लिपियों का एक परिवार है जिसे दुनिया भर के विभिन्न भाषाई समुदायों द्वारा व्यापक रूप से अपनाया और इस्तेमाल किया जाता है। ऐसी लिपियों को इस्तेमाल करने वाली विभिन्न भाषाओं की पहचान करना भाषा टेक्नोलॉजी (प्रौद्योगिकी) के लिए बहुत ज़रूरी है और कम-संसाधनीय विन्यासों (लो-रिसोर्स: ऐसी भाषाएँ जिनकी इंटरनेट पर मौजूदगी अभी अंग्रेजी, फ्रेंच, अरबी की तरह व्यापक नहीं है) में चुनौतीपूर्ण है। यह पेपर फ़ारसी-अरबी लिपियों का इस्तेमाल करने वाली भाषाओं की सटीक पहचान करने के रास्ते में चुनौतियों पर प्रकाश डालता है, विशेष रूप से द्विभाषी समुदायों में जहाँ भाषाओं को "अपरंपरागत" रूप से लिखा जाता है (यानी भाषा की खुद की लिपि के बजाये प्रभुत्वशाली भाषा की लिपि में लेखन)। इस समस्या के समाधान के रूप में, वाक्यों को भाषाओं में वर्गीकृत (क्लास्सिफ़ाई) करने के लिए हम पर्यवेक्षित तकनीकों (सुपरवाइज़्ड लर्निंग) का इस्तेमाल करते हैं। इसके आधार पर, हम एक ऐसे पदानुक्रम मॉडल (पदानुक्रम/हाइरार्किकल मॉडल कंप्यूटर विज्ञान की ऐसी तकनीक है जो किसी भी फ़ैसले को छोटे छोटे चरणों में तोड़ देती है और एक एक कर फ़ैसले लेती है) का प्रस्ताव रखते है जो उन भाषाओं के कलस्टर/समूह को निशाना बनाती है जो अक्सर मॉडल को उलझा देता है (ऐसी भाषाएं जिनमें मॉडल के लिए अंतर करना मुश्किल हो)। हमारे प्रयोग के नतीजे हमारे प्रस्ताव की प्रभावशीलता की ओर इशारा करते हैं।[1]

1 भूमिका (Introduction)

ऐतिहासिक रूप से, अरब फ़तह के क्षेत्रीय विस्तार के कारण दुनिया में कई दीर्घकालिक बदलाव हुए, विशेष रूप से नस्ली-भाषाई (एथ्नो-लिंगविस्टिक) दृष्टिकोण से, जहाँ उस समय की स्थानीय भाषाओं को अपने अस्तित्व को बनाये रखने के लिए चुनौतियों का सामना करना पड़ा (Wasserstein, 2003) । अरबी के प्रशासनिक भाषा - राइख़स्पराख़ यानी शाही भाषा- होने की वजह से कई स्थानीय भाषाएँ अपनी शब्दावली और लेखन में प्रभावित हुईं। कई सदियों तक, फ़ारसी ने अपनी ख़ास ध्वनियों के लिए नए लेखिम (ग्राफ़ीम अक्षर) जोड़कर अरबी लिपि का विस्तार किया। जैसे <پ> (<प>, U+067E) और <گ> (<ग>, U+06AF)□ इसलिए, शास्त्रीय अरबी लिपि के मुख्य विस्तारित रूपों में से एक फ़ारसी-अरबी लिपि है जिसे धीरे-धीरे कई अन्य भाषाओं द्वारा अपनाया गया है, मुख्य रूप से पश्चिम, मध्य और दक्षिण एशिया में (Khansir and Mozafari, 2014) । फ़ारसी-अरबी लिपि का इस्तेमाल करने वाली कुछ भाषाएँ हैं उर्दू, कुर्दी, पश्तो, अज़ेरी तुर्की, सिंधी और उइघुर। इन आधुनिक भाषाओं के अलावा ओटोमन तुर्की जैसी कई भाषाएँ हैं जिन्होंने ऐतिहासिक रूप से इस लिपि का इस्तेमाल किया है। हालाँकि ऐसी अन्य लिपियाँ भी हैं जिन्हे फ़ारसी-अरबी लिपि से प्रभावित हुए बिना सीधे अरबी लिपि से बनाया गया था जैसे कि कुछ अफ़्रीकी भाषाओं में इस्तेमाल की जाने वाली अजामी लिपि (जैसे स्वाहिली और वोलोफ़), दक्षिणी एशिया में इस्तेमाल की जाने वाली पेगॉन और जावी लिपियाँ, और ऐतिहासिक रूप से कुछ यूरोपीय भाषाओं के लिए इस्तेमाल की जाने वाली अल्जामियादो लिपि।

'भाषा पहचान' दस्तावेज़, वाक्य और उप-वाक्य जैसे विभिन्न स्तरों पर किसी पाठ/टेक्स्ट की भाषा का पता लगाने का कार्य है। मशीन अनुवाद और सूचना पुनर्प्राप्ति (इनफ़ॉर्मेशन रिट्रीवल) की तरह प्राकृतिक भाषा प्रसंस्करण (एनएलपी - नेचुरल लैंग्वेज प्रोसेसिंग) में इस कार्य के महत्व को देखते हुए, इसका बड़े पैमाने पर अध्ययन किया गया है और इसे भावना विश्लेषण (सेंटीमेंट एनालिसिस) और मशीन अनुवाद जैसे विभिन्न अनुप्रयोगों के लिए फायदेमंद दिखाया गया है (Jauhiainen et al., 2019) । यह कार्य सभी विन्यासों और भाषाओं के लिए समान रूप से चुनौतीपूर्ण नहीं है, क्योंकि यह दिखाया जा चूका है कि छोटे पाठों की या अति-सम्बंधित भाषाओं की पहचान (भाषाई तौर पर और लेखन में भी) बहुत मुश्किल है । मिसाल के तौर पर: फ़ारसी बनाम दारी, या कुर्दी के विभिन्न प्रकारों की पहचान (Malmasi et al., 2015; Zampieri et al., 2020) ।

इसके अलावा, द्विभाषी समुदायों में बोली जाने वाली कुछ कम-संसाधनीय भाषाओं को अपनी मातृभाषा के लिए प्रशासनिक या शैक्षणिक समर्थन की कमी या सीमित प्रोद्योगिकी के कारण लेखन में विभिन्न चुनौतियों का सामना करना पड़ता है। परिणामस्वरूप, पाठ अपरंपरागत रूप से लिखा जाने लगता है - यानी भाषा की पारंपरिक लिपि या वर्तनी के अनुसार नहीं, बल्कि प्रशासनिक रूप से "प्रमुख'' या "प्रभुत्वशील" भाषा की लिपि पर निर्भर होकर। मिसाल के तौर पर, कश्मीरी या कुर्दी को कभी-कभी उनकी ख़ुद की विस्तृत फ़ारसी-अरबी लेखनविधि (ऑर्थोग्राफ़ी) के बजाय, क्रमशः उर्दू या फ़ारसी लिपियों में लिखा जाता है। इससे उन भाषाओं की पहचान करना और भी जटिल हो जाता है, जिससे लिपि-

[1] डेटा और मॉडल https://github.com/sinaahmadi/PersoArabicLID पर उपलब्ध हैं|

भाषा	639-3	विपि	लिपि प्रकार	विशेषक	ZWNJ	प्रभुत्व
अज़ेरि तुर्की	azb	azb	अब्जद	✓	✓	फ़ारसी
गिलाकि	glk	glk	अब्जद	✓	✓	फ़ारसी
मज़ांदेरानी	mzn	mzn	अब्जद	✓	✓	फ़ारसी
पश्तो	pus	ps	अब्जद	✓	✗	फ़ारसी
गोरानी	hac	-	वर्णमाला	✗	✗	फ़ारसी, अरबी, सोरानी
उत्तरी कुर्दी (कुरमानजी)	kmr	-	वर्णमाला	✗	✗	फ़ारसी, अरबी
मध्य कुर्दी (सोरानी)	ckb	ckb	वर्णमाला	✗	✗	फ़ारसी, अरबी
दक्षिणी कुर्दी	sdh	-	वर्णमाला	✗	✗	फ़ारसी, अरबी
बलोची	bal	-	अब्जद	✓	✗	फ़ारसी, उर्दू
ब्राहुई	brh	-	अब्जद	✓	✗	उर्दू
कश्मीरी	kas	ks	वर्णमाला	✓	✗	उर्दू
सिंधी	snd	sd	अब्जद	✓	✗	उर्दू
सरैकी	skr	skr	अब्जद	✓	✗	उर्दू
तोरवाली	trw	-	अब्जद	✓	✗	उर्दू
पंजाबी	pnb	pnb	अब्जद	✓	✗	उर्दू
फ़ारसी	fas	fa	अब्जद	✓	✓	-
अरबी	arb	ar	अब्जद	✓	✗	-
उर्दू	urd	ur	अब्जद	✓	✓	-
उइघुर	uig	ug	वर्णमाला	✗	✗	-

Table 1: इस पेपर में चयनित भाषाओं की फ़ारसी-अरबी लिपियाँ। कॉलम 2 और 3 ISO 639-3 में और यदि उपलब्ध हो तो उनके विशिष्ट विकिपीडिया (विपि) पर भाषाओं के कोड दिखाते हैं। विशेषक (डायाक्रिटिक) और जीरो-विड्थ नॉन-जॉइनर (ZWNJ) कॉलम व्यक्तिगत वर्णों के रूप में विषेकों और ZWNJ के उपयोग को संदर्भित करते हैं।

यों की समानता के कारण भ्रम पैदा होते हैं और डेटा की कमी के कारण डेटा-आधारित तकनीकें बाधित होती है। इसलिए, फ़ारसी-अरबी लिपियों का उपयोग करने वाली भाषाओं की विश्वसनीय भाषा पहचान आज भी एक चुनौती बनी हुई है, ख़ास तौर से कम-संसाधनीय भाषाओं में।

इस प्रकार, हम कुछ ऐसी भाषाओं का चयन करते हैं जो फ़ारसी-अरबी लिपियों का उपयोग करती हैं, जिनका सारांश टेबल 1 में दिया गया है । इनमें से अधिकांश न केवल डेटा की कमी बल्कि अपरंपरागत लेखन से संबंधित चुनौतियों का सामना कर रहीं हैं। इसलिए, हम इन भाषाओं के लिए भाषा पहचान कार्य को दो विन्यासों में परिभाषित करते हैं - (क) पाठ अपनी भाषा की लिपि/वर्तनी के अनुसार लिखा जाता है, जिसे **पारंपरिक लेखन** कहा जाता है, या (ख) पाठ में प्रशासनिक रूप से प्रभावी भाषा की लिपि या शब्दावली के उपयोग के कारण कुछ हद तक विसंगतियाँ हैं, जिसे **अपरंपरागत लेखन** कहा जाता है। यह ध्यान में रखते हुए कि फ़ारसी-अरबी लिपियाँ ज़्यादातर पाकिस्तान, ईरान, अफ़ग़ानिस्तान और इराक़ की मूल भाषाओं को लिखने में इस्तेमाल की जाती हैं, हम उर्दू, फ़ारसी और अरबी को अपने शोध में शामिल करते हैं क्योंकि वे अपने क्षेत्रों की प्रशासनिक रूप से प्रभुत्वशाली भाषाएँ हैं। इसके अलावा, भाषाओं का एक विविध समुच्चय होने से यह भी पता चल सकता है कि कौन सी भाषाएँ अक्सर एक-दूसरे के रूप में भ्रमित होती हैं। बेशक़ हम उइघुर को भी शामिल करते हैं, यह ध्यान दिया जाना चाहिए कि यह मुख्य रूप से एक द्विभाषी समुदाय में बोली जाती है, यानी चीन के सिनच्यांग में, □□□□□ अपरंपरागत लेखन फ़ारसी-अरबी लिपि नहीं है; इसलिए, हम उइघुर के लिए केवल पारंपरिक लेखन पर विचार करते हैं।

योगदान यह पेपर फ़ारसी-अरबी लिपि या इसकी विभिन्न विस्तारित लिपियों में लिखी गई भाषाओं की भाषा पहचान पर प्रकाश डालता है। हम स्क्रिप्ट मैपिंग (प्रतिचित्रण या लिप्यांतरण) का इस्तेमाल करके डेटा एकत्र करने और संश्लेषणात्मक-शोरगुल (सिंथेटिक-नॉइज़) वाक्य बनाने का वर्णन करते हैं(§2) । हम कुछ वर्गीकरण तकनीकों को लागू (इम्प्लीमेंट) करते हैं और अतिसम्बंधित भाषाओं के बीच भ्रम को हल करने के लिए एक पदानुक्रम मॉडल तरीक़े का प्रस्ताव करते हैं। प्रस्तावित तकनीक अन्य तकनीकों से बेहतर प्रदर्शन करती है और शोरगुल विन्यास (सेटिंग) में 0.88-0.95 के बीच मैक्रो-औसत F_1 प्राप्त करती है (§3)।

ZWNJ - ज़ीरो-विड्थ नॉन-जॉइनर एक वर्चुअल ग़ैर-मुद्रण वर्ण है जिसका उपयोग डिजिटलीकृत लेखनविधियों में किया जाता है जिनमें संयुक्ताक्षरों का इस्तेमाल होता है। कुछ लिपियों में वर्ण, शब्द में उनके स्थान के आधार पर, रूप बदलते हैं और कुछ वर्णों के मिश्रण पर एक संयुक्ताक्षर बनता है। इससे यह मिलन नहीं होता और वर्णों को क्रमशः उनके अंतिम और प्रारंभिक रूपों में ही मुद्रित होते हैं।

ISO 639-3 भाषाओं के नामों के प्रतिनिधित्व के लिए तीन-अक्षरीय कोड हैं। इसका उद्देश्य सभी ज्ञात प्राकृतिक भाषाओं (जीवित एवं विलुप्त) को व्यापक रूप से कवर करना है।

2 पद्धति

चूँकि चुनी गई भाषाएँ ज़्यादातर कम-संसाधनीय हैं, उनके लिए डेटा इकट्ठा करना और परंपरागत या अपरंपरागत तरीक़ों से लिखे लेखों की पहचान करना बहुत दुर्जेय काम है। इस समस्या से निपटने के लिए, हम वेब/इंटरेंट पर मौजूद विभिन्न स्रोतों से डेटा इकट्ठा करने पर ध्यान देते हैं।[2]

फिर, हम संश्लेषणात्मक डेटा बनाने वाली एक ऐसी तकनीक का प्रस्ताव रखते हैं जो संभावित रूप से अपरंपरागत लेखन में मौजूद विभिन्न प्रकार के शोर (नॉइज़) को प्रतिबिंबित (रिफ़्लेक्ट) और मॉडल कर सकती है। इसलिए, हम एक सरल तकनीक का इस्तेमाल करते हैं जो एक भाषा की लिपि के अक्षरों को दूसरी (प्रभुत्वशाली) भाषा की लिपि में मैप करती है। और आख़िर में, हम इस टास्क/कार्य को बेंचमार्क करने के अपने प्रयासों की चर्चा करते हैं और एक ऐसे पदानुक्रम मॉडल का प्रस्ताव रखते हैं जो संबंधित भाषाओं के बीच मॉडल की उलझन को सुलझता है।

2.1 डेटा संग्रहण/कलेक्श

जैसा कि टेबल 1 में दर्शाया गया है, गोरानी, उत्तरी एवं दक्षिणी कुर्दी, बलूची, ब्राहुई, और तोरवाली के अलावा सभी भाषाओं के उनकी फ़ारसी-अरबी लिपि इस्तेमाल करने वाले समर्पित विकिपीडिया पन्ने हैं। इसीलिए, हम उपलब्ध भाषाओं के लिए विकिपीडिया डंप्स[3] को कॉर्पोरा (कोष) के रूप में इस्तेमाल करते हैं। वहीं दूसरी ओर, उत्तरी एवं दक्षिणी कुर्दी, बलूची और ब्राहुई के लिये, हम टेबल A.2 में दी गयीं न्यूज़ वेबसाइटों को क्रॉल[4] कर डेटा इकट्ठा करते हैं। इसके अलावा, हम Uddin and Uddin (2019) का तोरवाली कार्पस/कोष,Ahmadi (2020) का गोरानी कार्पस, Esmaili et al. (2013) का मध्य कुर्दी का कॉर्पस, और Tehseen et al. (2022) का पंजाबी कार्पस। फ़ारसी, अरबी, और उर्दू के लिए हम ततोएबा (Tatoeba) डेटासेट का प्रयोग करते हैं।[5]

डेटा को एकत्रित करने के बाद, हम विभिन्न फॉर्मेट को मूल (रॉ) लेख में बदलते हैं, लेख का पूर्वप्रसंस्करण करते हैं (प्रीप्रोसेसिंग), फ़ॉर्मेटिंग शैलियों से संबंधित विशेष अक्षरों/वर्णों को हटाने के लिए रेग्युलर ऍक्सप्रैशन (नियमित व्यंजक) का उपयोग करते हैं, और ईमेल, फ़ोन नंबर, वेबसाइट यूआरएल संबंधित जानकारी हटा देते हैं। हम सभी प्रस्तुत अंकों को लैटिन अंकों में बदलते हैं क्योंकि आम तौर पर फ़ारसी-अरबी लेखों में अंकों के एक मिश्रण का इस्तेमाल होता है। इस मिश्रण में फ़ारसी अंक <۰۱۲۳۴۵۶۷۸۹> , अरबी अंक <٠١٢٣٤٥٦٧٨٩> और लैटिन अंक अलग अलग अनुपात में मौजूद होते हैं। यह बदलाव यह सुनिश्चित करता है कि भाषा पहचान टास्क के लिये वाक्यों में बाद में विविधि प्रकार के अंकों को संश्लेषणात्मक रूप से जोड़ा जा सके। चूँकि कुछ चुनी भाषाएँ दो लिपियों का इस्तेमाल करती हैं, जैसे कि पंजाबी की गुरमुखी और शाहमुखी लिपियाँ, और कश्मीरी की देवनागरी और फ़ारसी-अरबी लिपियाँ, हमने कॉर्पोरा में मौजूद लिपि और कोड स्विच्ड[6] या कोटेड वाक्यों को हटाने के लिए कुछ नियमित व्यंजकों का प्रयोग भी किया। चूँकि ऐसे मिश्रित वाक्यों की पहचान करना ही काफ़ी मुश्किल है, हम यह स्पष्ट करना चाहते हैं कि साफ़-कॉर्पोरा में भी कुछ कोड स्विच्ड शब्द हो सकते है।

हम वर्णों की यूनिकोड एनकोडिंग को एकीकृत कर लेख पूर्वप्रसंस्करण को अंजाम देते हैं। यूनिकोड एन्कोडिंग[7] में विसंगतियां अक्सर अलग अलग बाइंडिंग वाले कीबोर्ड के उपयोग की वजह से होती हैं और पहले ही कुछ भाषाओं के पूर्वप्रसंस्करण चरण में शामिल हैं। (Ahmadi, 2019; Doctor et al., 2022)□ उदाहरण के तौर पे, <ے> (U+06D2) और <ي> (U+064A) को <ی> (U+06CC) के बजाए और <ك> को कुर्दी में(U+0643) <ک> (U+06A9) के बजाए इस्तेमाल किया जा सकता है। ज़ीरो-विड्थ नॉन-जॉइनर (ZWNJ, U+200C) के टेबल 1 में दर्शाए गए इस्तेमाल के अनुसार, हम इसे पूर्वप्रसंस्करण चरण में शामिल करने पर भी विचार करते हैं।[8] और अंत में, हम रेगुलर व्यंजकों का इस्तेमाल कर कॉर्पोरा को वाक्यों के स्तर पर टोकनाइज़ करते हैं।

टेबल A.3 में एकत्रित कॉर्पोरा में मौजूद १० सबसे ज़्यादा बारंबार ट्राइग्राम्स को प्रस्तुत किया गया है,जिनमें यह देखा जा सकता है कि कई संयोजक (कंजंक्शन) और प्रत्यय (एफ़िक्स) भी पुनर्प्राप्त किए गये हैं जो भाषा की पहचान का इशारा बन सकते हैं।

2.2 लिपि बदलाव/मैपिंग

यह मान कर कि शोरगुल लेख हमेशा प्रभुत्वशाली भाषा की लिपि या वर्तनी में लिखे होंगे, हम किसी दी गई भाषा की फ़ारसी-अरबी लिपि को प्रभुत्वशाली लिपि में बदलते हैं। उदाहरण: कश्मीरी लिपि को उर्दू लिपि में, मध्य कुर्दी लिपि को फ़ारसी और अरबी लिपियों में बदलना। यह करने के लिए, हम वर्णों की यूनिकोड एनकोडिंग और उनके बीच दृश्य समानता पर निर्भर करते हैं:

- यदि एक लेखिम दोनों भाषाओं की लिपियों में मौजूद हैं, जैसे कि सिंधी और उर्दू में <ھ> (U+06BE) या सरैकी और उर्दू में <ٹ> (U+0679), तो हम उन्हें एक साथ मैप करते हैं, इस बात पर ध्यान दिये बग़ैर कि दोनों भाषाओं में इस वर्ण का उच्चारण कैसे होता हैं।

[2] https://www.wikidata.org

[3] **डंप** - किसी तारीख़ को विकिपीडिया के सभी पृष्ठों पर मौजूद डेटा को एक जगह एक फ़ाइल में इकठ्ठा/डाउनलोड करने को उस तारीख़ का विकिपीडिआ डंप कहेंगे। हमारे इस प्रयोग में हम 20 जनवरी 2023 के डंप का इस्तेमाल करते हैं।

[4] **क्रॉलर** या वेब स्पाइडर एक एल्गोरिदम/बॉट है जो व्यवस्थित रूप से इंटरनेट ब्राउज़ करता है और जिसे आमतौर पर किसी भी वेबपेज पर मौजूद डेटा को आँकने का और उसपर निगरानी रखने का काम दिया जाता है।

[5] https://tatoeba.org

[6] **कोड स्विचिंग** बहुभाषी लोगों द्वारा बोलते, लिखते, या संकेत देते समय दो भाषाओं के बीच अदल बदल करने को कहते हैं। इससे हिंग्लिश या तंगलिश जैसी मिश्रित भाषिकाओं का जन्म होता है। इसी प्रकार, लिपि स्विचिंग का अर्थ है कि एक ही वाक्य में दो लिपियों का प्रयोग।

[7] **एन्कोडिंग** का सरल अर्थ है किसी भी वस्तु, संख्या, वर्ण, या जानकारी को किसी अन्य प्रारूप में सहेजना। जैसे कि कीबोर्ड पर मौजूद अक्षर कंप्यूटर के अंदर अंकों के रूप में ही सहेजे जाते हैं और हर अक्षर के लिए एक अनूठे अंक का चयन किया जाता है।

[8] हम चुनी हुई भाषाओं में सामान्य/व्यापक लेखन पद्धतियों के बारे में जानकारी हासिल करने के लिए वेब पर विभिन्न स्रोतों की सलाह लेते हैं, ख़ास तौर पर https://scriptsource.org.

- यदि प्रभुत्वशाली लिपि में एक समरूप लेखिम मौजूद नहीं है, तो सबसे अधिक देखने में समान वर्ण को स्रोत वर्ण में मैप किया जाता है। उदाहरण: ब्राहुई के <ڵ> (U+06B7) से सबसे ज़्यादा मिलने वाला वर्ण उर्दू में <ل> (U+0644) है। गिलाकी के <ۆ> (U+06CB) को फ़ारसी के मिलने वाले <و> (U+0648) में मैप किया जाता है। इस तरह, एक वर्ण को स्रोत भाषा में कई वर्णों में मैप किया जा सकता है।
- कुछ मैपिंग लेखनविधि नियमों का पालन करती हैं, ख़ास तौर पर उन वर्णों के लिए जो शब्द में उनकी जगह के अनुसार शकल बदलते हैं। उदाहरण: कुर्दी स्वर एक आरंभिक हमज़ा <ئ> (U+0626) के साथ दिखते हैं, जैसे <ئو> /oː/ और <ئ> /eː/। हम ऐसे नियमों को भी शामिल करते हैं।
- चूँकि अंकों को डेटा संग्रहण चरण में एकीकृत कर लिया गया है(§2.1), हम लैटिन अंकों को फ़ारसी और अरबी अंकों में बेतरतीब (रैंडम) मैप भी करते हैं।

प्रभुत्वशाली भाषाओं के अनुसार, हर स्रोत और प्रभुत्वशाली भाषा के लिए हम एक लिपि मैपिंग मैन्युअल रूप से तैयार करते हैं। यह ध्यान में रखना चाहिये कि ग़ैर-विशेषक वर्णों के साथ साथ, विशेषक वर्णों को भी शामिल किया गया है अगर वह विशेषक (जैसे हरकत) एक लेखिम का भाग हैं, जैसे कि गोरानी और सिंधी में <ڏ> (U+068E), पर <ا> नहीं। अलग हो सकने वाली 'हरकात' जैसे फ़त्हा, कसरा, दम्मा लिपि-मैपिंग में शामिल नहीं की गई हैं। टेबल A.1 चयनित भाषाओं में उनके अरबी, फ़ारसी, और उर्दू (फ़ारसी-अरबी लिपि उपयोग करने वाली तीन प्रमुख भाषाएँ) से रिश्ते के अनुसार इस्तेमाल किए गए वर्णों के समुच्चय को प्रस्तुत करता है।

2.3 संश्लेषणात्मक डेटा उत्पादन (सिंथेटिक डेटा जनरेशन)

लिपि मैपिंग का इस्तेमाल कर, हम 'साफ़' वाक्यों के आधार पर संश्लेषणात्मक वाक्यों को बनाकर, अपरंपरागत लेखन की नक़ल करते हैं, यानी संग्रहित कॉर्पोरा में मौजूद वाक्य। 'साफ़' वाक्य में वर्णों को एक टारगेट लिपि के विकल्प में बेतरतीब बदलकर इस कार्य को अंजाम दिया जाता है (मैपिंग का इस्तेमाल करके)। शोर/नॉइज़ के भाषा पहचान पर असर का मूल्यांकन करने के लिए, हम अलग अलग शोर के स्तर पर (20% से लेकर 100% तक) डेटा का संश्लेषण (सिंथेसाइज़) करते हैं, जहाँ संभावित प्रतिस्थापनों (सब्स्टिट्यूशंज़) के आधार पर शोर को लागू किया गया है। टेबल 2 उत्तरी कुर्दी में एक साफ़ वाक्य और शोर के स्तरों के आधार पर उसके संश्लेषणात्मक-शोरगुल समकक्ष वाक्यों को दिखाती है। इसीलिए, सभी दटसेट को ऐसे श्रेणीबद्ध (कैटेगोराइज़) किया गया है:

1. **क्लीन/साफ़**: डेटासेट जिसमें बिना किसी शोर इंजेक्ट किए कॉर्पोरा के मूल वाक्य हैं। यह डेटा में 0% शोर के समान है। इस विन्यास में चयनित कम-संसाधनीय भाषाएँ, उर्दू, फ़ारसी, अरबी, और उयघुर सभी शामिल हैं।
2. **शोरगुल/नोइज़ी**: डेटासेट जिसमें 20% से लेकर 100% तक के शोरगुल वर्णों वाले वाक्य शामिल हैं। चलन पर ध्यान न देते हुए, जब शोर 100% हो, अलग हो सकने वाले विशेषकों को हटाया गया है, कश्मीरी सहित जिसमें विशेषक हर लफ़्ज़ पर लगाने अनिवार्य हैं। हम सभी शोर स्तरों को एक कर एक नये डेटासेट का निर्माण भी करते हैं - **सभी/ऑल**। चूँकि फ़ारसी, अरबी,उर्दू, और उयघुर अपरंपरागत लेखन का सामना नहीं करते, उन्हें शोरगुल डेटा में शामिल नहीं किया गया है।
3. **मर्ज्ड**:साफ़/क्लीन और सभी/ऑल डेटा को मर्ज करने का परिणाम

Noise %	Sentence
Clean	دووەمین پێشانگەها فۆتۆگرافەرێن کورد ل بەلجیکا Second Kurdish photographers' exhibition in Belgium
20	دووهمین پێشانگهها فۆتۆگرافهرێن کورد ل بهلجیکا
40	دووه مین بشانکه ها فطکرافه رن کورد ل به لجیکا
60	دووة مین بشانکة ها فوتوکرافة رن کورد ل بة لجیکا
80	دووەمین بیشانکەها فۆتۆکرافەرین کورد ل بەلجیکا
100	دووهمین بیشانکهها فوتوکرافهرین کورد ل بهلجیکا

Table 2: उत्तरी कुर्दी (कुर्मांजी) का एक वाक्य, उसके अलग अलग शोर के स्तर के अनुसार संश्लेषणात्मक रूप से बनाए वाक्य)

3. **मर्ज्ड**:साफ़/क्लीन और सभी/ऑल डेटा को मर्ज करने का परिणाम

साफ़ और **शोरगुल** डेटासेट में मिलाकल हर भाषा के लिए 10,000 वाक्य हैं, ब्राहुई, तोरवाली, और बलूची को छोड़ कर जिनके लिए क्रमशः सिर्फ़ 549, 1371, और 1649 वाक्य कॉर्पोरा में उपलब्ध हैं। इसलिए, इन 3 भाषाओं से हम 500 वाक्यों को परीक्षण-सेट में डालकर बचे हुए वाक्यों को 4 के गुणक (कोएफ़िशिएंट) के साथ अपसैंपल करते हैं, यानी शेष वाक्यों की 4 गुना नक़ल करना और उनको प्रशिक्षण-सेट मानें। इसी तरह, कश्मीरी और गोरानी जिनके लिए क्रमशः 6340 और 8742 वाक्य उपलब्ध हैं, 2000 वाक्य पहले परीक्षण-सेट में जोड़े जाते हैं, और शेष वाक्यों को अपसैंपल किया जाता जिससे हमें प्रशिक्षण-सेट के लिए 8000 वाक्य मिलते हैं। असंतुलन से बचने के लिए, प्रभुत्वशाली भाषाएँ जिनके डेटा में कोई शोर नहीं है, जैसे उर्दू, फ़ारसी, अरबी, उयघुर, उनमें 10,000 नये इंस्टैंस/वाक्य उनके साफ़ कॉर्पोरा से जोड़े गये हैं। तो, मर्ज्ड डेटा में हर भाषा के लिए 20,000 साफ़/क्लीन और शोरगुल वाक्य हैं ।

2.4 बेंचमार्किंग

हम भाषा पहचान पर एक प्रयिक्तात्मक (प्रोबाबिलिस्टिक) वर्गीकरण समस्या की तरह विचार करते हैं, जिसमें हर वाक्य की एक विशिष्ट वर्ग (जैसे भाषा) में होने की प्रागुक्ति (प्रेडिक्शन) की जाती है। हम पिछले भागों में बताये गये विभिन्न डेटासेट के प्रशिक्षण-परीक्षण सेट के वाक्यों के 80/20 विभाजन का इस्तेमाल करते हैं। दोनों ही सेट एक डेटा से बनाये गये हैं।

मूलाधार/बेसलाइन सिस्टम के रूप में, हम फ़ास्टटेक्स्ट का पूर्वप्रशिक्षित भाषा पहचान मॉडल`lid.176` इस्तेमाल

करते हैं, जो कि विकिपीडिया, ततोएबा (Tatoeba) , और सेटाइम्ज़ (SETimes) के 176 भाषाओं के डेटा पर प्रशिक्षित किया गया है, जिनमें सभी चयनित भाषाएँ शामिल हैं, बलूची, ब्राहुई, गिलाकी, गोरानी, उत्तरी कुर्दी (फ़ारसी-अरबी लिपि में), दक्षिणी कुर्दी, और तोरवाली के अलावा। इसके साथ, हम फ़ास्टटेक्स्ट के साथ, 64 साइज़ के वर्ड-वेक्टर, 2-6 के करैक्टर एन-ग्रैम्स की न्यूनतम और अधिकतम लंबाई/लेंथ, 1.0 का लर्निंग रेट (सीखने की दर), 25 इपोक, और एक पदानुक्रम सॉफ्टमैक्स व्यय (लॉस) के साथ एक मॉडल प्रशिक्षित करते हैं।

फ़ास्टटेक्स्ट-संबंधित मूलाधार और हमारे ख़ुद के मॉडलों के अलावा, हम गूगल के CLD3 (Salcianu et al., 2020), फ्रैंक (Franc)[9], और लैंगआईडी.पाई(Langid.py) (Lui and Baldwin, 2012) जैसी नवीनतम (स्टेट ऑफ़ द आर्ट) तकनीकों की बेंचमार्किंग के लिये परिशुद्धता, प्रत्याह्वान , और एफ़-1 स्कोर भी रिपोर्ट करते हैं। हम दो और मूलाधार/बेसलाइन शेयर करते हैं जो 2-4 साइज़ के करैक्टर एन-ग्राम, मल्टीनोमियल नाइव बेज़ मॉडल (नॉन-यूनिफ़ॉर्म्ड लर्नड क्लास प्रायर, बिना लपलेस स्मूदिंग), और अधिकतम 500 इटरेशन, 500 साइज़ की एक हिडन लेयर, और 1000 बैच साइज़ के मल्टीलेयर पर्सेप्ट्रोन, के साथ शून्य से प्रशिक्षित हैं।

2.5 पदानुक्रम मॉडलिंग

पदानुक्रम मॉडलिंग (फ़िगर 1) का लक्ष्य है अतिसंबंधित भाषाओं के बीच भ्रम को मिटाना, जिसे वो ऐसे एक्सपर्ट वर्गीकारक (क्लासीफायर) प्रशिक्षित करके पूरा करता है जो कुछ चुनिंदा भाषाओं के बीच पहचान करने में माहिर होते हैं। यह करने के लिए, हम बेस्ट-परफॉरमेंस वाले मॉडल (प्रशिक्षण डेटा पर) की कन्फ़्यूशन-मैट्रिक्स का निरीक्षण करते हैं और ऐसे भाषाई क्लस्टर पहचानते हैं जिनमें मॉडल काफ़ी ज़्यादा भ्रम दर्शाता है। पिछले भाग में उल्लिखित कस्टम-प्रशिक्षित फ़ास्टटेक्स्ट मॉडल रूट-वर्गीकारक का रोल निभाता है, और हम उसकी कन्फ़्यूशन-मैट्रिक्स से तीन क्लस्टर पहचानते हैं: (फ़िगर 2):

1. **क्लस्टर 1**: उत्तरी, दक्षिणी, मध्य कुर्दी, और गोरानी
2. **क्लस्टर 2**: फ़ारसी, गिलाकी, मज़ानदेरानी, अज़ेरी तुर्की, पश्तो
3. **क्लस्टर 3**: उर्दू, कश्मीरी, पंजाबी, सिंधी, सरैकी

पदानुक्रम पेड़ की हर उपइकाई फ़ास्टटेक्स्ट मॉडल है जो उचित क्लस्टर की भाषाओं के डेटा पर शून्य से प्रशिक्षित है (रूट मॉडल की तरह समान पैरामीटर)।

3 नतीजे

टेबल 3 में, हम सभी डेटासेट, 6 SOTA और अनुकूलित-प्रशिक्षित बेसलाइन, हमारे रूट फास्टटेक्स्ट मॉडल (रूट), और एक पदानुक्रमित भ्रम-समाधान मॉडल (हायर) में परिशुद्धता, प्रत्याह्वान और F_1 स्कोर का विवरण करते हैं। हमने पाया कि हमारा रूट फास्टटेक्स्ट मॉडल पूर्व-प्रशिक्षित फास्टटेक्स्ट बेसलाइन, गूगल के CLD3, लैंगएआईडी.पाई, फ्रैंक, और एमएलपी की तुलना में काफी मार्जिन से अच्छा प्रदर्शन करता है।

3.1 अत्याधुनिक बनाम सरल बेसलाइन

तीन अत्याधुनिक मॉडल (CLD3, `langid.py`, `Franc`) में से किसी को भी सभी 19 भाषाओं और शोरगुल विन्यासों में हमारे परीक्षण सेट पर 0.15 F_1 से अधिक स्कोर नहीं मिला। असल में, मिश्रित-शोरगुल विन्यासों (40% - सभी) के लिए उन्हें अक्सर बेहद कम F_1 स्कोर ($0 \leq F_1 < 0.1$) मिलते हैं। यह इन मॉडलों के उर्दू, फ़ारसी, अरबी, सिंधी के सपोर्ट के बावजूद है, जिसमें फ्रैंक अतिरिक्त रूप से मध्य कुर्दी को भी कवर करता है। यह सैकड़ों भाषाओं को कवर करने के दावों के बावजूद अत्याधुनिक पूर्व-प्रशिक्षित मॉडलों में भाषा पहचान की ख़राब गुणवत्ता को दर्शाता है, और यह भी बताता है कि भाषा पहचान अभी तक पूर्ण रूप से 'हल' नहीं हुई है। इन तीन मॉडलों की तुलना में, एमएनबी और एमएलपी मॉडल सभी शोरगुल स्तरों (20% शोर को छोड़कर) में बेहतर प्रदर्शन करते हैं, और यहां तक कि 8 में से 7 शोर विन्यासों पर फास्टटेक्स्ट के बड़े पूर्व-प्रशिक्षित मॉडल `lid.176` से भी बेहतर प्रदर्शन करते हैं। इससे वह `lid.176` मॉडल की तुलना में अधिक बेहतर मॉडल साबित होता है।

3.2 फ़ास्टटेक्स्ट के साथ पदानुक्रमित मॉडलिंग

अब हमारे दो मॉडलों, अनुकूलित फास्टटेक्स्ट मॉडल (रूट) और पदानुक्रमित भ्रम-रिज़ॉल्यूशन मॉडल (हायर) पर नज़र डालते हैं। यह साफ़ दिखता है कि दोनों मॉडल बड़े अंतर से किसी भी बेसलाइन मॉडल की तुलना में बेहतर प्रदर्शन करते हैं। चूंकि पदानुक्रमित मॉडल को मिश्रित डेटासेट पर प्रशिक्षित किया जाता है, जिसमें स्वच्छ (0% शोरगुल) विन्यास की तुलना में चार अधिक वर्गों के साथ शोरगुल और साफ़ वाक्य शामिल हैं, इसलिए यह स्वाभाविक है कि रूट मॉडल स्वच्छ विन्यास में बेहतर प्रदर्शन करता है। हालाँकि, किसी भी व्यावहारिक शोरगुल स्तर (20% से मर्ज्ड तक) के लिए पदानुक्रमित मॉडल रूट मॉडल से बेहतर प्रदर्शन करता है।

इन सूक्ष्म सुधारों का परीक्षण करने के लिए, हम एक-पूंछ वाले ज़ेड टेस्ट (वन टेल्ड ज़ेड टेस्ट) के अनुसार प्रत्येक शोर स्तर के लिए सांख्यिकीय महत्व रखने वाले नतीजों को रिपोर्ट करते हैं। इससे हम जड़ मॉडल की तुलना पदानुक्रमित मॉडल के साथ, महत्व स्तर 0.01 पर करते हैं। हम ज़ेड-परीक्षण करते हैं क्योंकि नमूनों की संख्या 30 से अधिक है और नमूना भिन्नता को जनसंख्या भिन्नता के अनुमान के रूप में विश्वसनीय रूप से उपयोग किया जा सकता है। शून्य परिकल्पना (नल हाइपोथेसिस) यह है कि रूट और पदानुक्रमित मॉडल ($\mu_0 : f_{root} = f_{hier}$) के बीच कोई महत्वपूर्ण अंतर नहीं है और वैकल्पिक परिकल्पना (अल्टरनेट हाइपोथिसिस) है कि पदानुक्रमित मॉडल का प्रदर्शन रूट मॉडल की तुलना में महत्वपूर्ण और सख़्ती रूप से अधिक है ($\mu_1 : f_{root} < f_{hier}$)। हम रूट मॉडल के F_1 स्कोर f_{root} के लिए एक-वन टेल्ड 99% विश्वास अंतराल की गणना करते हैं। एक-पूंछ वाले ज़ेड-टेस्ट के अनुसार, हम शून्य परिकल्पना

[9] https://github.com/wooorm/franc/

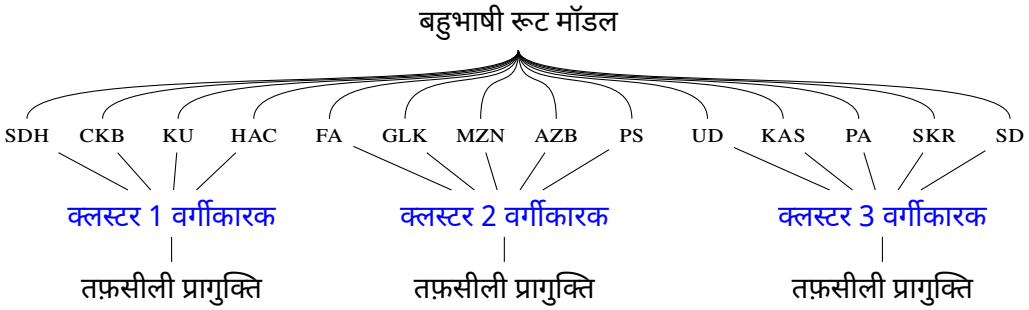

Figure 1: पदानुक्रम मॉडल की वास्तुकला/आर्किटेक्चर। यदि रूट मॉडल दक्षिणी कुर्दी(SDH), गोरानी(HAC) , उत्तरी कुर्दी(KMR) ,या मध्य कुर्दी(CKB) की प्रागुक्ति करता है, तो सैंपल वाक्य को निचले स्तर पर मौजूद एक एक्सपर्ट वर्गीकारक को भेजा जाता है जो इन चार भाषाओं के बीच भेद करने में प्रशिक्षित है। क्लस्टर 2 और 3 के लिए भी ऐसे ही। यदि रूट मॉडल द्वारा किसी अनक्लस्टर्ड भाषा की प्रागुक्ति की जाती है (यानी कोई भी शाखाएँ उपलब्ध नहीं हैं), तो पदानुक्रम मॉडल भी रूट मॉडल वाली भाषा की प्रागुक्ति करेगा।

को अस्वीकार कर सकते हैं और निष्कर्ष निकाल सकते हैं कि F_1 स्कोर के बीच का अंतर सांख्यिकीय रूप से महत्वपूर्ण है यदि पदानुक्रमित मॉडल का F_1 स्कोर f_{hier} इस अंतराल की ऊपरी सीमा के अंदर सख़्ती से खत्म हो जाता है।

टेबल 4 में, हम अपनी परिकल्पना परीक्षण के नतीजों को रिपोर्ट करते हैं और पाते हैं कि हमारे पदानुक्रमित भ्रम-समाधान दृष्टिकोण द्वारा मिला लाभ सभी शोरगुल विन्यासों के लिए 99% आत्मविश्वास स्तर पर सांख्यिकीय रूप से महत्वपूर्ण है। इसलिए, हम स्थापित करते हैं कि पूरे मॉडल को फिर से प्रशिक्षित किए बिना शोरगुल डेटा पर प्रदर्शन में सुधार करने के लिए एक भ्रम-आधारित पदानुक्रमित मॉडल का उपयोग किया जा सकता है और यह सांख्यिकीय रूप से परीक्षण सेट में महत्वपूर्ण सुधार लाता है ।

3.3 भाषा-विशेष प्रदर्शन

टेबल 5 में, हम दो सर्वश्रेष्ठ मॉडल के लिए सभी शोरगुल विन्यासों के लिए भाषा के स्तर पर स्कोर रिपोर्ट करते हैं: हमारा अनुकूलित फास्टटेक्स्ट मॉडल और हमारा भ्रम-रिज़ॉल्यूशन पदानुक्रमित मॉडल। सभी भाषाओं और शोरगुल विन्यासों में, पदानुक्रमित मॉडल 128 विन्यासों में से केवल 5 में ख़राब प्रदर्शन करता है। अन्य सभी के लिए, यह या तो रूट मॉडल के बराबर या उससे बेहतर प्रदर्शन करता है। बोल्ड में दिये गये आँकड़े दर्शाते हैं कि पदानुक्रमित मॉडल तीनों भ्रम समूहों में शोरगुल विन्यासों (20%-सभी) में सबसे अधिक सुधार लाता है।

जैसा कि अपेक्षित था, उन भाषाओं के लिए जो किसी अत्यधिक भ्रमित क्लस्टर का हिस्सा नहीं थीं, यानी अरबी, बलूची, तोरवाली, उयघुर और ब्राहुई, पदानुक्रमित और रूट मॉडल समान प्रगुक्तियाँ उत्पन्न करता है, इसलिए, शोरगुल स्तरों पर समान स्कोर दिखायी देता है। टेबल A.4 में, हम अपने मॉडल की तुलना में पूर्व-प्रशिक्षित फास्टटेक्स्ट मॉडल की प्रगुक्तियों के आधार पर विभिन्न शोरगुल स्तरों पर कुछ भाषा पहचान उदाहरण भी प्रस्तुत करते हैं।

4 संबंधित शोध

मॉडलिंग के तरीक़े भाषा पहचान को आम तौर पर एक बहु-वर्ग पाठ वर्गीकरण कार्य के रूप में तैयार किया जाता है और इसने सभी भाषाओं और उपभाषिकाओं और सीमित डेटा विन्यासों में सीधे बाइट, करैक्टर या शब्द-स्तर n-ग्राम सुविधाओं के साथ अत्याधुनिक प्रदर्शन हासिल किया है (Jauhiainen et al., 2017) । मॉडल या वर्गीकारक का चुनाव स्रोत, प्रक्षेत्र (डोमेन) और प्रति भाषा डेटा की मात्रा पर अत्यधिक निर्भर है, सपोर्ट वेक्टर मशीन (Ciobanu et al., 2018; Malmasi and Dras, 2015) और मल्टीनोमियल नाइव बेयस (King et al., 2014; Mathur et al., 2017) जैसे सरल रैखिक वर्गीकारक सीमित डेटा के साथ मज़बूत बेसलाइन प्रदान करते हैं और सभी डोमेन में गणना करते हैं। यदि बड़ी मात्रा में डेटा उपलब्ध है, तो एकत्रित वर्गीकारक (Baimukan et al., 2022) और तंत्रिका (न्यूरल) मॉडल का उपयोग किया जा सकता है, लेकिन यह ध्यान रहे कि समान भाषा किस्मों और बोलियों के साथ इन मॉडलों को फिर भी दिक़्क़त आएगी और ओवरफिटिंग (Medvedeva et al., 2017; Criscuolo and Aluísio, 2017; Eldesouki et al., 2016) का ख़तरा रहेगा। इस पेपर में, हम भाषा पहचान के लिए एक पदानुक्रमित मॉडल का प्रस्ताव करते हैं जो शोरगुल विन्यासों में आम तौर पर भ्रमित होने वाली भाषाओं के बीच का फ़र्क़ कर सकेगा और छोटी वर्गीकरण इकाइयों के साथ ऐसी ग़लत प्रगुक्तियों को हल कर सकेगा। इस तरह के मॉडल का उपयोग भाषा कवरेज का विस्तार करने और बड़े और कंप्यूटिंग-भूखे मॉडल को फिर से प्रशिक्षित किए बिना मौजूदा पूर्व-प्रशिक्षित मॉडल के प्रदर्शन में सुधार करने के लिए किया जा सकता है। हमारे मामले में, हमने शोरगुल डेटा विन्यासों के लिए सांख्यिकीय रूप से महत्वपूर्ण सुधार देखे।

मिलती जुलती भाषाएँ और भाषिकाएँ भाषा पहचान पर बहुत अध्ययन किया जा चुका है, और इसीलिए इसे कभी-कभी हल किया हुआ भी माना जाता है; वास्तव में, विश्व की अधिकांश भाषाएँ वर्तमान प्रणालियों द्वारा समर्थित/सपोर्टेड नहीं हैं। प्रतिनिधित्व की यह कमी बड़े पैमाने पर डेटा खनन प्र-

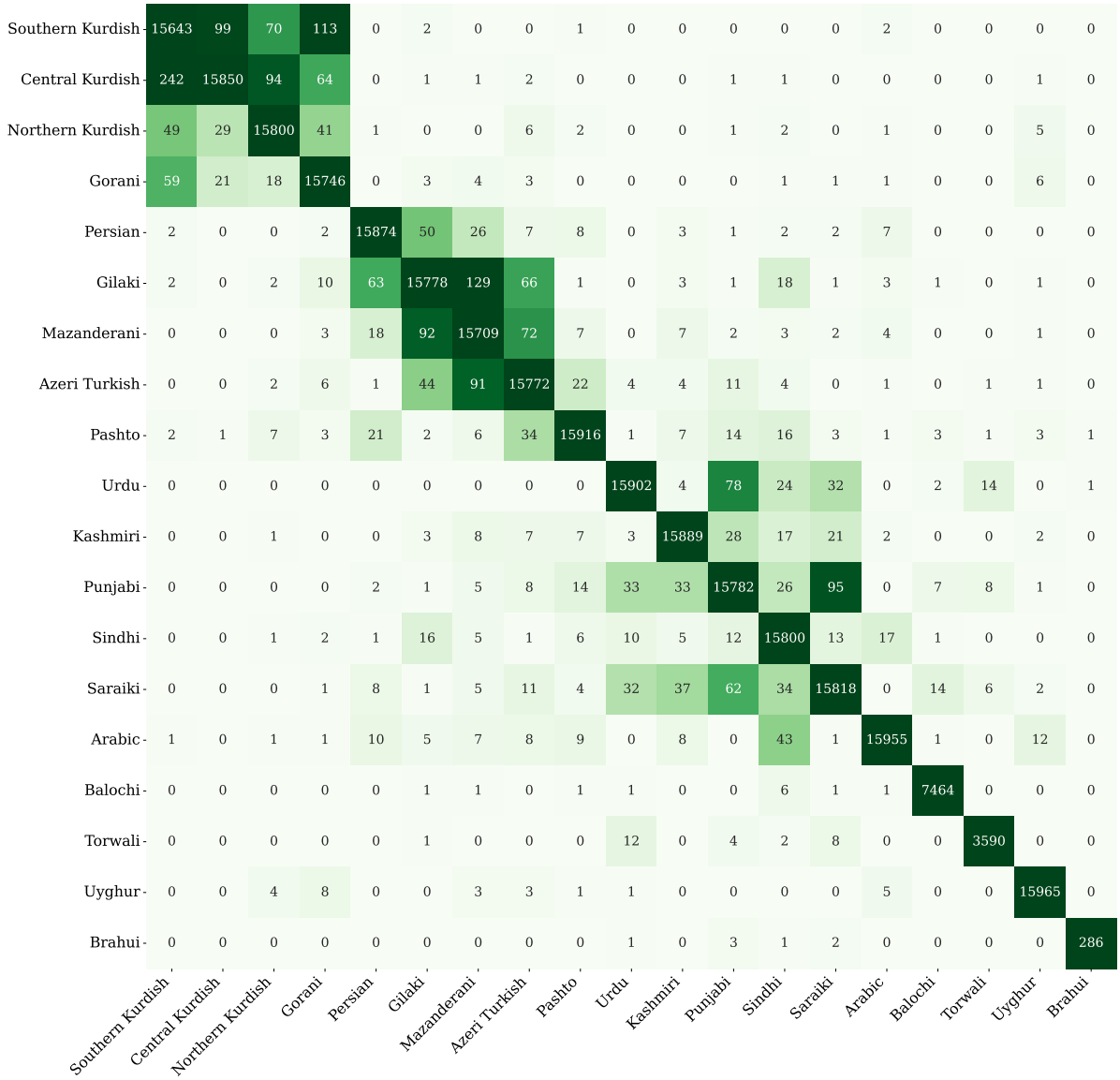

Figure 2: प्रशिक्षण डेटासेट पर बहुभाषी रूट मॉडल का भ्रम मैट्रिक्स। पंक्ति लेबल हमारे अनुकूलित फास्टटेक्स्ट मॉडल की प्रगुक्तियों को दर्शाते हैं, कॉलम गोल्ड लेबल (प्रशिक्षण डेटासेट) को दर्शाते हैं, और प्रत्येक सेल एक (प्रगुक्ति, गोल्ड लेबल) जोड़ी के लिए मॉडल द्वारा की गई प्रगुक्तियों की संख्या को दर्शाती है। भ्रम मैट्रिक्स से, हमने तीन अत्यधिक भ्रमित भाषा समूहों की पहचान की, जैसा कि अनुभाग 2.5 में बताया गया है

शोर	माप	पदानुक्रम	रूट	फ़ास्टटेक्स्ट	CLD3	langid.py	Franc	MNB	MLP
0%	परिशुद्धता	0.72	**0.91**	0.16	0.03	0	0.02	0.43	0.47
	प्रत्याह्वान	0.70	**0.89**	0.07	0.05	0	0.02	0.14	0.16
	F_1 स्कोर	0.72	**0.90**	0.10	0.04	0	0.02	0.21	0.24
20%	परिशुद्धता	**0.92**	0.92	0.30	0.08	0.13	0.13	0.08	0.03
	प्रत्याह्वान	**0.89**	0.89	0.32	0.18	0.18	0.18	0.05	0.05
	F_1 स्कोर	**0.91**	0.90	0.31	0.11	0.15	0.15	0.06	0.04
40%	परिशुद्धता	**0.91**	0.90	0.17	0.04	0	0.01	0.51	0.49
	प्रत्याह्वान	**0.88**	0.88	0.07	0.05	0	0	0.09	0.11
	F_1 स्कोर	**0.90**	0.89	0.10	0.05	0	0	0.16	0.19
60%	परिशुद्धता	**0.91**	0.90	0.17	0.04	0	0	0.45	0.54
	प्रत्याह्वान	**0.88**	0.87	0.07	0.05	0	0	0.12	0.09
	F_1 स्कोर	**0.89**	0.88	0.09	0.04	0	0	0.20	0.15
80%	परिशुद्धता	**0.90**	0.90	0.16	0.03	0	0	0.25	0.33
	प्रत्याह्वान	**0.88**	0.87	0.06	0.05	0	0	0.12	0.15
	F_1 स्कोर	**0.89**	0.88	0.08	0.04	0	0	0.16	0.21
100%	परिशुद्धता	**0.90**	0.90	0.15	0.03	0	0	0.44	0.44
	प्रत्याह्वान	**0.88**	0.87	0.06	0.05	0	0	0.08	0.11
	F_1 स्कोर	**0.89**	0.88	0.08	0.03	0	0	0.13	0.17
सभी	परिशुद्धता	**0.90**	0.89	0.15	0.03	0	0	0.28	0.51
	प्रत्याह्वान	**0.87**	0.86	0.06	0.05	0	0	0.16	0.10
	F_1 स्कोर	**0.88**	0.88	0.08	0.04	0	0	0.20	0.17
मर्ज्ड	परिशुद्धता	**0.95**	0.95	0.28	0.06	0.11	0.11	0.15	0.15
	प्रत्याह्वान	**0.94**	0.94	0.27	0.16	0.16	0.16	0.08	0.07
	F_1 स्कोर	**0.95**	0.94	0.27	0.09	0.13	0.13	0.10	0.10

Table 3: सभी शोर विन्यासों के लिए सारे भाषाई पहचान मॉडलों की परिशुद्धता, प्रत्याह्वान, और F_1 स्कोर। हमारा पदानुक्रम और रूट मॉडल सभी शोर विन्यासों के लिये सबसे अच्छे दो मॉडल हैं। फ़ास्टटेक्स्ट, मल्टीनोमियल नाइव बेज़, और मल्टीलेयर पर्सेप्ट्रोन अलग अलग शोर विन्यासों के लिये तीसरे स्थान पर आते हैं। परिशुद्धता, प्रत्याह्वान, और F_1 स्कोर सभी बेंचमार्कों के लिये पेश किए गए हैं। यदि कोई दो ऐसे अंक हैं जो दशमलव के सौवें स्थान तक एक से हैं, तो बोल्ड करके बेहतर अंक को दर्शाया गया है।

शोर	टेस्ट प्रतिदर्श	Δ	सार्थक
0	33500	-0.188	✗
20	25500	0.005	✓
40	25500	0.006	✓
60	25500	0.007	✓
80	25500	0.007	✓
100	25500	0.007	✓
सभी	27806	0.007	✓
मर्ज्ड	69304	0.002	✓

Table 4: सभी शोर विन्यासों के लिए हमारा पदानुक्रम मॉडलिंग द्वारा रूट मॉडल की तुलना में F_1 स्कोर में पाये गये सभी सुधार सांख्यिकी रूप से सार्थक हैं (सार्थकता स्तर=0.01, यानी 99% विश्वास्यता)

यासों को प्रभावित करती है और कम संसाधन वाली भाषाओं के लिए डेटा की कमी को और बढ़ा देती है। भाषा पहचान प्रणालियों में भाषा कवरेज को बेहतर बनाने में एक प्रमुख बाधा समान भाषाओं, भाषा किस्मों और बोलियों के बीच अंतर करने की क्षमता है। जैसा कि इस पेपर में बताया गया है, यह तब और भी चुनौतीपूर्ण हो जाता है जब एक भाषा समुदाय एक प्रमुख भाषा की अपरंपरागत लिपि को अपनाता है। हाल ही में, नॉर्डिक भाषाओं (Haas and Derczynski, 2021) , अरबी बोलियों (Nayel et al., 2021; Abdul-Mageed et al., 2020; Salameh et al., 2018) और क्षेत्रीय इतालवी और फ्रेंच भाषा किस्मों (Jauhiainen et al., 2022; Camposampiero et al., 2022) के बीच अंतर करने के लिए अध्ययन हुए हैं। उदाहरण के लिए, Haas and Derczynski (2021) , छह नॉर्डिक भाषाओं: डेनिश, स्वीडिश, नॉर्वेजियन (निनोर्स्क), नॉर्वेजियन (बोकमाल), फ़रोईज़ और आइसलैंडिक के बीच सर्वोत्तम अंतर करने के लिए कई मॉडलिंग और फीचराइजेशन मॉडल के साथ प्रयोग करते हैं।

	0%		20%		40%		60%		80%		100%		सभी		मर्ज्ड	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
समूह 1																
SDH	0.95	**0.96**	0.95	**0.96**	0.94	**0.95**	0.93	**0.94**	0.93	**0.94**	0.94	0.94	0.94	0.94	0.95	**0.96**
CKB	0.95	0.95	0.94	0.94	0.92	**0.94**	0.91	**0.93**	0.91	**0.93**	0.91	**0.92**	0.92	**0.93**	0.95	0.95
KU	0.95	0.95	0.93	**0.94**	0.93	0.93	0.92	**0.93**	**0.93**	0.92	0.92	0.92	0.92	**0.93**	0.95	0.95
HAC	0.94	0.94	0.94	0.94	0.93	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.91	**0.92**	0.94	0.94
समूह 2																
FA	0.97	**0.98**	-	-	-	-	-	-	-	-	-	-	-	-	0.97	**0.98**
GLK	0.92	**0.94**	0.88	**0.89**	0.88	**0.89**	0.88	**0.9**	0.88	**0.89**	0.88	**0.89**	0.92	0.92	0.92	**0.94**
MZN	0.92	0.92	0.85	**0.86**	0.85	**0.86**	0.85	**0.87**	0.85	**0.86**	0.85	**0.87**	0.92	**0.93**	0.92	0.92
AZB	0.91	0.91	0.86	**0.87**	0.85	**0.86**	0.86	**0.87**	0.86	**0.87**	0.85	**0.86**	0.9	**0.91**	0.91	0.91
PS	0.96	0.96	0.94	**0.95**	0.94	**0.95**	0.94	**0.95**	0.94	**0.95**	0.94	0.94	0.96	0.96	0.96	0.96
समूह 3																
UD	0.96	**0.97**	-	-	-	-	-	-	-	-	-	-	-	-	0.96	**0.97**
KAS	0.94	**0.95**	0.9	**0.91**	0.9	**0.91**	0.9	**0.91**	0.9	**0.91**	0.87	**0.88**	**0.91**	0.9	0.94	**0.95**
PA	0.91	0.91	**0.87**	0.86	0.86	0.86	0.86	0.86	0.85	**0.86**	0.85	0.85	0.87	0.87	0.91	0.91
SD	0.93	**0.94**	0.89	**0.91**	0.88	**0.89**	0.87	**0.89**	0.87	**0.89**	0.87	**0.89**	0.91	0.91	0.93	**0.94**
SKR	**0.92**	0.91	0.85	0.85	0.84	**0.85**	0.84	**0.85**	0.85	0.85	0.84	**0.85**	0.86	**0.88**	**0.92**	0.91
AR	0.98	0.98	-	-	-	-	-	-	-	-	-	-	-	-	0.98	0.98
BAL	0.98	0.98	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.97	0.97	0.98	0.98
TRW	0.95	0.95	0.87	0.87	0.89	0.89	0.88	0.88	0.88	0.88	0.87	0.87	0.91	0.91	0.95	0.95
UG	0.99	0.99	-	-	-	-	-	-	-	-	-	-	-	-	0.99	0.99
BRH	0.84	0.84	0.7	0.7	0.67	0.67	0.68	0.68	0.68	0.68	0.65	0.65	0.63	0.63	0.84	0.84

Table 5: हमारे रूट (M1) और पदानुक्रम (M2) मॉडलों के लिये हर भाषा के F_1 स्कोर। हमारा पदानुक्रम मॉडल तीनों समूहों में (और अलग अलग शोर विन्यासों में) F_1 स्कोर में सुधार दर्शाता है। डैश के साथ दिये सेल दिखाते हैं कि उस भाषा की अंदर केवल पारंपरिक लेखन की रीत है और इसीलिए उसे किसी भी संश्लेषणात्मक डेटा विन्यासों का भाग नहीं बनाया गया।

उन्होंने पाया कि फास्टटेक्स्ट से निकाले गए स्किपग्राम एम्बेडिंग समृद्ध हैं और निकट-संबंधित भाषाओं के बीच अंतर करने में सक्षम हैं। यह ध्यान देने योग्य है कि जबकि पेपर ने चयनित भाषाओं में सुधार प्रस्तुत किया है, सभी छह चयनित नॉर्डिक भाषाओं में बड़ी मात्रा में प्रशिक्षण डेटा (50K+ वाक्य) हैं और वे पहले से ही लैंगआईडी.पाई जैसे ऑफ-द-शेल्फ टूल द्वारा समर्थित हैं। समान भाषाओं और बोलियों के बीच अंतर करने के लिए, नाइव बेज़ और लॉजिस्टिक रिग्रेशन जैसे अधिक उथले और रैखिक वर्गीकारक, एमएलपी या कन्वेन्शनल तंत्रिका नेटवर्क जैसे तंत्रिका मॉडल से बेहतर प्रदर्शन करते हैं (Chakravarthi et al., 2021; Aepli et al., 2022; Ceolin, 2021) । इसकी पुष्टि गैर-तंत्रिका शास्त्रीय मशीन लर्निंग दृष्टिकोण द्वारा की जाती है, जो कि द्रविड़ भाषाओं, रोमानियाई बोलियों, इतालवी और फ्रेंच क्षेत्रीय किस्मों जैसी विशिष्ट रूप से विविध भाषाओं में (Jauhiainen et al., 2022; Camposampiero et al., 2022) और यूरेलिक भाषाओं में (Chakravarthi et al., 2021) वारडायल 2021 और 2022 के अधिकांश शेयर्ड टास्क को जीतता है। तंत्रिका मॉडलिंग दृष्टिकोण, समान भाषाओं/किस्मों में सीमित डेटा के कारण, कभी-कभी गैर-तंत्रिका आधार रेखाओं का प्रदर्शन भी कम कर सकते हैं जैसा कि यूरालिक भाषा पहचान या इतालवी बोली पहचान शेयर्ड टास्कों में बताया गया है (Chakravarthi et al., 2021; Aepli et al., 2022) ।

5 निष्कर्ष

हम अपने अध्ययन को द्विभाषी समुदायों में लिखी जाने वाली भाषाओं पर केंद्रित करते हैं जहाँ एक पारंपरिक और बेहतर फ़ारसी-अरबी लिपि संस्करण के स्थान पर एक अपरंपरागत प्रभुत्वशाली फ़ारसी-अरबी लिपि का उपयोग किया जाता है। हम डेटा संकलन और भाषा पहचान दोनों में इस परिस्थिति की अनोखी चुनौतियों पर चर्चा करते हैं, और ऐसी अपरंपरागत लेखन विधियों से सामना होने पर SOTA प्रणालियों में परिणामस्वरूप प्रदर्शन-संबंधित मुद्दों पर चर्चा करते हैं। इस मुद्दे पर शोरगुल और साफ़/मिश्रित विन्यासों के बीच F_1 स्कोर में 20-पॉइंट के अंतर द्वारा रोशनी डाली गई है।

हमारा प्रस्तावित पदानुक्रमित मॉडल एक अनुकूलित-प्रशिक्षित फास्टटेक्स्ट सिस्टम, सरल MNB (मल्टीनोमियल नाइव बेज़) और MLP (मल्टी लेयर पर्सेप्ट्रोन) और Google के CLD3, "फ्रैंक" और "लैंगआईडी.पाई" की SOTA भाषा पहचान प्रणालियों से बेहतर प्रदर्शन करता है। हम रूट बहुभाषी मॉडल के भ्रम मैट्रिक्स का विश्लेषण करने के बाद एक पदानुक्रमित मॉडल का उपयोग करके सांख्यिकीय रूप से महत्वपूर्ण सुधार पाते हैं।

6 सीमाएँ

चयनित भाषाओं में से कुछ भाषाएँ एक से ज़्यादा लिपि का प्रयोग करती हैं - जैसे कि पंजाबी या कुर्दी। इससे एकत्रित किए डेटा की गुणवत्ता कमतर हो सकती है क्योंकि इसे आम

तौर पर स्वचालित रूप से पूर्वप्रसंस्कृत किया जाता है। इस वजह से हमारा यह मानना है कि हमारे डेटासेट में कोड-स्विच्ड डेटा एक नगण्य मात्रा में मौजूद है। और तो और, फ़ारसी-अरबी लिपियों पर ध्यान केंद्रित करने के कारण, हमने ऐसी भाषाओं की अन्य लिपियों के पाठों को इस शोध में शामिल नहीं किया। हालाँकि एक भाषा एक से अधिक प्रभुत्वशाली भाषाओं से प्रभावित हो सकती है और संश्लेषणात्मक डेटा विभिन्न लिप्यंतरणों से उत्पन्न होता है, इस शोध में हमने हर प्रभुत्वशाली भाषा के विशेष असर का विश्लेषण नहीं किया। इस दिशा को अंजाम देने के लिए, हर प्रभुत्वशाली भाषा के लिए एक अधिकतम बारीक वर्गीकरण टास्क का इजात होना चाहिए। इसके अलावा, फ़ारसी की दारी और फ़ारसी जैसी भाषिकाओं और अन्य चयनित भाषाओं की उपभाषिकाओं को इस शोधकार्य में शामिल किया जा सकता है। इसी तरह, हमारी पदानुक्रमित तकनीक को अन्य लिपियों पर भी लागू किया जा सकता है, ख़ास तौर से उन लिपियों पर जिन्हें कई भाषाओं द्वारा इस्तेमाल किया जाता है, जैसे कि सिरिलिक और लैटिन। अंत में, हमारे द्वारा एकत्रित इस डेटा के आधार पर अन्य तकनीकों को लागू और फाइनट्यून किया जा सकता है।

आम तौर पर यह माना जाता है कि प्रस्तुत मॉडल जितने ज़्यादा डेटा पर प्रशिक्षित होंगे, उतने ही बेहतर होते जाएँगे । हम यह स्वीकार करते हैं कि रूट मॉडल की तुलना में हमारे पदानुक्रमित मॉडल के सुधार हमारे प्रशिक्षण सेट की मात्रा/साइज़ से सीमित हैं। अधिक वास्तविक शोरगुल डेटा प्राप्त होने पर, यह संभव है कि सभी शोरगुल सेटअपों के साथ-साथ साफ़ डेटा सेटअप में भी हमारा प्रदर्शन बेहतर होगा।

आभार

इस कार्य को DEL/DLI पुरस्कार BCS-2109578 के तहत राष्ट्रीय विज्ञान फाउंडेशन (NSF) द्वारा और पुरस्कार PR-276810-21 के तहत 'राष्ट्रीय मानविकी अक्षय निधि' द्वारा उदारतापूर्वक समर्थन दिया गया था। लेखक बनाम समीक्षकों के साथ-साथ जॉर्ज मेसन विश्वविद्यालय (GMU) में अनुसंधान कंप्यूटिंग कार्यालय (ORC) के भी आभारी हैं, जहाँ सभी कम्प्यूटेशनल प्रयोग आयोजित किए गए थे।

उद्धरण

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Sina Ahmadi. 2019. A rule-based Kurdish text transliteration system. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–8.

Sina Ahmadi. 2020. Building a Corpus for the Zaza-Gorani Language Family. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2020, Barcelona, Spain (Online), December 13, 2020*, pages 70–78. International Committee on Computational Linguistics (ICCL).

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596.

Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. The curious case of logistic regression for Italian languages and dialects identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Andrea Ceolin. 2021. Comparing the performance of CNNs and shallow models for language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–112, Kiyv, Ukraine. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, and Liviu P. Dinu. 2018. Discriminating between Indo-Aryan languages using SVM ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 178–184, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcelo Criscuolo and Sandra Maria Aluísio. 2017. Discriminating between similar languages with word-level convolutional neural networks. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 124–130, Valencia, Spain. Association for Computational Linguistics.

Raiomond Doctor, Alexander Gutkin, Cibu Johny, Brian Roark, and Richard Sproat. 2022. Graphemic Normalization of the Perso-Arabic Script. *arXiv preprint arXiv:2210.12273*.

Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish. 2016. QCRI @ DSL 2016: Spoken Arabic dialect identification using textual features. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 221–226, Osaka, Japan. The COLING 2016 Organizing Committee.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for Sorani Kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

René Haas and Leon Derczynski. 2021. Discriminating between similar Nordic languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. Italian language and dialect identification and regional French variety detection using adaptive naive Bayes. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–129, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluation of language identification methods using 285 languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 183–191, Gothenburg, Sweden. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *J. Artif. Intell. Res.*, 65:675–782.

Ali Akbar Khansir and Nasrin Mozafari. 2014. The impact of Persian language on Indian languages. *Theory and Practice in Language Studies*, 4(11):2360.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015. Language identification using classifier ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 35–43, Hissar, Bulgaria. Association for Computational Linguistics.

Shervin Malmasi, Mark Dras, et al. 2015. Automatic language identification for Persian and Dari texts. In *Proceedings of PACLING*, pages 59–64.

Priyank Mathur, Arkajyoti Misra, and Emrah Budur. 2017. LIDE: language identification from text documents. *CoRR*, abs/1701.03682.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 2017. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 156–163, Valencia, Spain. Association for Computational Linguistics.

Hamada Nayel, Ahmed Hassan, Mahmoud Sobhi, and Ahmed El-Sawy. 2021. Machine learning-based approach for Arabic dialect identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 287–290, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *27th International Conference on Computational Linguistics, COLING 2018*, pages 1332–1344. Association for Computational Linguistics (ACL).

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2020. Compact Language Detector v3 (CLD3).

Amina Tehseen, Toqeer Ehsan, Hannan Bin Liaqat, Amjad Ali, and Ala Al-Fuqaha. 2022. Neural POS tagging of Shahmukhi by using contextualized word representations. *Journal of King Saud University-Computer and Information Sciences*.

Naeem Uddin and Jalal Uddin. 2019. A step towards Torwali machine translation: an analysis of morphosyntactic challenges in a low-resource language. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 6–10, Dublin, Ireland. European Association for Machine Translation.

David J Wasserstein. 2003. Why did Arabic succeed where Greek failed? Language change in the Near East after Muhammad. *Scripta Classica Israelica*, 22:257–272.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

A चयनित भाषाएँ

Graphemes	Kurdish	Gorani	Uyghur	Arabic	Azeri	Gilaki	Mazanderani	Persian	Urdu	Kashmiri	Punjabi	Saraiki	Torwali	Pashto	Sindhi	Brahui	Balochi
Arabic	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی ح خ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی ح خ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی خ غ ف ق ي ك	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق ة ي ك	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ ث ذ ح خ ص ض ط ظ ع غ ف ق	ئ ا ب ت ج د ر ز س ش ل م ن و ه ی آ أ إ ؤ
Persian	پ ک چ گ ژ	پ ک چ گ ژ	پ چ گ ژ		پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ ژ	پ ک چ گ	پ ک چ گ ژ	پ ک چ گ ژ
Urdu	ھ	ھ	ھ						ھ ے ٹ ں ڑ ڈ	ھ ے ٹ ں ڑ ڈ	ھ ے ٹ ں ڑ ڈ	ھ ے ٹ ں ڑ ڈ	ھ ے ٹ ں ڑ ڈ	ے ں	ھ	ھ ے ٹ ں ڑ ڈ	ھ ے ٹ ڈ
Language-specific	ڕ ە ڤ ڵ ۆ ێ ۊ	ڕ ە ڤ ڵ ۆ ێ ۊ ۋ ۆ ێ ۉ ڎ	ڭ ە ۇ ۆ ۈ ۋ ې		ۉ ۼ ۇ ۆ	ˇ ۊ ۆ	ˇ			ٲ ٳ ۄ ۆ ۆ ۓ	ݪ ڻ ݨ	ٻ ڄ ݙ ڳ ݨ	ځ ڇ څ ڂ ڗ ݜ	ځ څ ډ ړ ږ ټ ښ ګ ڼ ۀ ي ې ۍ	ڀ ٻ ٿ ٺ ٽ ڃ ڇ ڄ ڊ ڌ ڍ ڏ ڙ ڦ ڪ ڳ ڱ ڻ ي ۽ ۾	ڵ	ێ ۏ ˇ
Total/Unique	35/7	40/12	34/7	35/0	41/4	40/3	38/1	37/0	43/0	50/6	46/3	48/5	49/8	52/13	58/21	44/1	32/3

Table A.1: अरबी, फ़ारसी और उर्दू लिपियों के तुलनात्मक अवलोकन के साथ चयनित भाषाओं में प्रयुक्त फ़ारसी-अरबी लिपियाँ। ध्यान दें कि भाषा-विशेष के वर्ण उन वर्णों को संदर्भित करते हैं जो किसी भाषा के लिए अनोखे हैं और अरबी, फ़ारसी या उर्दू में उपयोग नहीं किए जाते हैं। इसे अंतिम पंक्ति में भी दर्शाया गया है।

भाषा	वेबसाइट
बलूची	`https://sunnionline.us/balochi/`
ब्राहुई	`https://talarbrahui.com`
उत्तरी कुर्दी	`https://badinan.org`
दक्षिणी कुर्दी	`https://shafaq.com/ku`

Table A.2: स्थानीय समाचार वेबसाइटें जिनसे एकत्रित डेटा क्रॉल किया गया है

KMR	CKB	SDH	HAC	UIG	ARB	AZB	GLK	MZN	FAS	URD	KAS	PNB	SKR	TRW	PUS	SND	BRH	BAL
␣و␣	␣نی	␣له	␣و␣	␣ئا	␣ال	␣ای	␣ان	␣ون	␣ست	␣یں	␣چھ	␣دے	␣اں	␣سی	␣د␣	␣جي	␣نا	␣نت
␣ێن	␣له	␣له	␣به	نىڭ	␣أن	␣یر	ستا	␣نه	␣می	␣ہے	␣اَ	␣اں	␣دے	سی␣	␣په	جي␣	نا␣	␣اں
␣ان	انی	ئه␣	␣نه	␣ىڭ	␣لى	یند	␣ای	␣و␣	است	␣می	مَن	دے␣	␣یں	می␣	په␣	␣ان	␣ان	␣ان
هه␣	␣ان	وه␣	␣ان	نىن	␣من	␣ین	تان	␣تا	␣را	میں	چھِ	␣وں	␣وں	␣می	او␣	␣۾␣	اس␣	انت
ئه␣	به␣	␣ان	جه␣	␣نى	الم	␣ده	␣ته	و␣ه	␣را	ہیں	␣ھِ	␣تے	␣تے	مھ␣	␣او	␣ۓ␣	␣ٹی	␣نی
کو␣	␣ەی	␣و␣	␣ەی	لى	␣في␣	␣او	␣␣␣	ا␣ه	␣اس	␣کی	␣ہٕ	␣دی	␣دی	کی␣	د␣ه	جو␣	اٹ␣	␣که
␣یا	ئه␣	هیل	که␣	␣ان	␣في	␣ان	ا␣ه	␣ره	␣به	␣ہے	␣کھ	␣دی	دے␣	␣ھی	کې␣	ج␣ن	او␣	ان␣
␣ل␣	هوه	␣یل	␣جه	␣ئى␣	␣سا	ا␣ی	ایس	␣ای	␣ند	ا␣ہ	␣مَ	␣وچ	␣دا	م␣ی	␣له	␣آه	␣آ␣	انی
␣نا	␣و␣	کرد	انی	ى␣ئ	␣مي	␣و␣	وست	␣له	␣به	␣ہو	اَک	␣نے	␣تے	تھ␣	د␣ې	␣ھن	ا␣ن	که␣
دا␣	ئ␣ی	␣یگ	ئا␣	رى	␣من	␣دا	␣سه	␣ته	می_	␣نے	کھ	␣وچ	␣دی	بھ␣	␣کې	␣جو	دا␣	␣ێں

Table A.3: चयनित भाषाओं के एकत्रित किए गए कॉर्पोरा में सब से ज़्यादा बार आने वाले 10 त्रिग्राम। ␣ और _, क्रमशः, स्पेस और ZWNJ को दर्शाते हैं। सभी त्रिग्रामों के बीच, कई संयोजकों और प्रत्ययों को देखा जा सकता है, जैसे कि उत्तरी कुर्दी (KMR) में ␣و␣ (‘और’), और गोरानी (HAC) में␣که␣ (‘कि’)

Language	Noise %	Prediction (@1)		Sentence
		`lid.176`	Our’s	
Punjabi	0	Urdu	Azeri	اور لادینیت واشتراکیت کو جمہوریت کے حسین لبادہ میں پیش کردیا گیا ۔
Saraiki	0	Punjabi	Saraiki	کہیں وی زبان وادب تے تحقیق زیادہ تر کیفیتی
Sindhi	0	Sindhi	Sindhi	گهٽا دفعا هڪ عورت ساٿياڻي جنهن سان ڪوئي افلاطوني
Balochi	0	Urdu	Balochi	آیانی راہا کہ تئی مہر بوتگ اَنت گنجّ گْوار
Azeri	0	Persian	Azeri	قوزئی و دوغو سوریه موختار ایداره‌ائتمه‌سی
Gilaki	0	Persian	Gilaki	شوراب ایسم ایته روستا ایسه جه راستوپی دهستان
Persian	0	Persian	Persian	جوانی زمان فرا گرفتن دانایی است. پیری زمان تمرین کردن آن است.
Uyghur	0	Uyghur	Uyghur	ھەيدەكچىلىك تەرتىپىنى ئاياغلاشتۇرۇش توغرىسىدا كېسىم چىقىرىدۇ
Southern Kurdish	0	Sorani	Southern Kurdish	ڤایرۆس کۆرۆنا لەرێ دادوەر و پاریزەرەیل دەوام لە دادگای هەولێر وسان
Kashmiri	20	Urdu	Kashmiri	سودھا رانی چھ اَکھ ہِندوستانُے اَداکارہ یوس فِلمَن مَنٛز چھِ کامُ کَران.
Kashmiri	100	Urdu	Kashmiri	سودھا رانی چھ اکھ ہندوستانے اداکارہ یوس فلمن منز چھ کام کران.
Sorani	20	Persian	Sorani	ریژهی دهرجوانی ئهمسال له سالی پێشتر زیاتره
Sorani	100	Arabic	Sorani	ريژهي دهرجاني ئهمصال له صالي پيشطر زياطره

Table A.4: चयनित भाषाओं में कुछ उदाहरण और उनके साथ फ़ास्टटेक्स्ट के पूर्वप्रशिक्षित मॉडल (`lid.176`) की प्रगुक्तियाँ। इसके साथ, तुलना करने लिए संग्रहित डेटा पर हमारे एक प्रशिक्षित मॉडल का नतीजा भी दिखाया गया है