| No. | Language | Sarvam | KCC Agent’s Interface |
| --- | --- | --- | --- |
| 1 | Hindi | प्रश्न जुआड़ में दाना इतना हल्का और कम क्यों होता है? उत्तर खेत में नमी की कमी और उर्वरक कम पड़ने से वजह से यह होता रहता है। आप एक काम कीजिए फूल आते टाइम यूरिया को पानी में घोलकर छिड़कते रहिए। | प्रश्न: झुआर में दाना इतना हल्का और कम क्यों होता है? उत्तर: खेत में नमी की कमी और उर्वरक कम पड़ने से वजह से ये होता रहता है। आप एक काम कीजिए। फूल आते टाइम यूरिया को पानी में घोल कर छिड़कते रहिए। |
| 2 | Punjabi | ਮੀਕਾ ਸਿੰਘ ਨਾਲ ਉਹ ਚੁੰਮਣ ਵਾਦ ਵਿਵਾਦ ਵਿੱਚ ਫਸੀ ਰਹੀ। | ਮੀਕਾ ਸਿੰਘ ਨਾਲੋਂ ਚੁੰਮਣਵਾਦ ਵਿਵਾਦ' ਚ ਫਸੀ ਰਹੀ. |
| 3 | English | With this transition to the big time, the band shortened their name to Stay Below. | With this transition to the big time, the band shortened their name to stay below. |
| 4 | Malayalam | പതിമുഖം. \nപതിമുഖത്തെ സാധാരണമായി സപ്പാൻ മരൻ അല്ലെങ്കിൽ കിഴക്കൻ ഇന്ത്യൻ ചുവന്ന മരം എന്നാണ് വിളിക്കുന്നത്. \nതമിഴ്നാട്, കേരളം, കർണാടക, ആന്ധ്രപ്രദേശ്, പശ്ചിമബംഗാൾ എന്നീ സംസ്ഥാനങ്ങളിൽ കൃഷി ചെയ്യുന്ന ഒരു ഔഷധ വൃക്ഷമാണിത്. | പതിമുഖം പതിമുഖത്തെ സാധാരണമായി സപ്പാൻ മരം അല്ലെങ്കിൽ കിഴക്കൻ ഇന്ത്യൻ ചുവന്ന മരം എന്നാണ് വിളിക്കുന്നത്. തമിഴ്നാട്, കേരളം, കർണാടക, ആന്ധ്രാപ്രദേശ്, പശ്ചിമബംഗാൾ എന്നീ സംസ്ഥാനങ്ങളിൽ കൃഷി ചെയ്യുന്ന ഒരു ഔഷധവൃക്ഷമാണിത്. |
| 5 | Tamil | அற முறுக்கினால் அற்றுப் போகும் | அற முறுக்கினால் அற்று போகும். |
| 6 | Kannada | ಯಾಕಂದ್ರೆ ಅದು ಎಕ್ಸಾಮ್ ನಾವು ಕಂಪ್ಲೀಟ್ ಮಾಡ್ಕೊಂಡ್ರೆ ಮಾತ್ರ ನಾವು ಆ ಸೆಕ್ಷನ್ ಮುಗಿಸಿಕೊಳ್ಳೋಕೆ ಆಗೋದು. ಮತ್ತೆ ವಯಸ್ಸಾಗಿರೋರು ಇರ್ತಾರಲ್ವಾ ಮನೇಲಿ ಅವರು ನ್ಯೂಸ್ ಗಳನ್ನ ನೋಡೋದು, ವಾರ್ತೆಗಳನ್ನ ಮತ್ತೆ ಮತ್ತೆ | ಯಾಕಂದರೆ ಅದು ಎಕ್ಸಾಮ್ ನಾವು ಕಂಪ್ಲೀಟ್ ಮಾಡ್ಕೊಂಡ್ರೆ ಮಾತ್ರ ನಾವು ಆ ಸೆಕ್ಷನ್ನ ಮುಗಿಸ್ಕೊಳೋಕ್ಕಾಗೋದು. ಮತ್ತು ವಯಸ್ಸಾಗಿರೋ ಇರ್ತಾರಲ್ವ ಮನೇಲಿ ಅವರು ನ್ಯೂಸ್ಗಳನ್ನ ನೋಡೋದು, ವಾರ್ತೆಗಳನ್ನ ಅನ್ ಮತ್ತೆ ಮತ್ತು. |
| 7 | Telugu | బ్రతికినన్ని నాళ్ళు ఫలము విచ్చుట కాదు. | బ్రతికినని నాళ్లు ఫలము విచ్చుటగాదు. |
| 8 | Gujarati | ઉપરાંત દંતકથા પ્રમાણે તેમણે તમિલ ભાષાની રચના કરી છે. | ઉપરાંત દંતકથા પ્રમાણે તેમણે તમિલ ભાષાની રચના કરી છે. |
| 9 | Marathi | उंदीर खेळताना आनंदाने हसण्याचा आवाज काढतात. | उंदीर खेळताना आनंदाने हसण्याचा आवाज काढतात। |
| 10 | Odia | ସେ ଏକୋଇଶିରୁ ଅଧିକ ବର୍ଷ ଧରି ଏ ଦାୟିତ୍ୱ ନିର୍ବାହ କରିଥିଲେ। | ସେ ଏକୋଇଶରୁ ଅଧିକ ବର୍ଷ ଧରି ଏ ଦାୟିତ୍ୱ ନିର୍ବାହ କରିଥିଲେ। |
| 11 | Bengali | পদার্থবিজ্ঞানের নিয়মগুলো অন্তরজ সমীকরণ হিসেবে লেখা হয়। | পদার্থবিজ্ঞানের নিয়মগুলো অন্তরজ সমীকরণ হিসেবে লেখা হয়। |
| 12 | Assamese | Not Supported | সেই শাৰীকেইটাৰ আকাৰৰ বস্তু সিমানখিনি দূৰত্বৰ পৰা দেখিবলৈ বন্ধ কৰি পেলালে। |
| 13 | Bodo | Not Supported |  |
| 14 | Dogri | Not Supported | मुड़े दे कुड़ी दे टिबड़े मलोंदे ते जि'यां टिबड़ा उं'दा मिलदा फिर ओह्दे बाद कुड़ी मुड़ा करदे मलाटी |
| 15 | Kashmiri | Not Supported | تُہۍ کَتہِ پٮ۪ٹھ چھِو |
| 16 | Konkani | Not Supported | बरें बरें म्हळ्यार दुदाचें तुमचें जातलें दूद म्हशींचें जाल्यार सादारण साठ रुपया लिटर आसा |
| 17 | Maithili | Not Supported | बहुत ही नीक लागैए बहुत अच्छा लागैए खाना बनाबैमे खाना पीना बनेलहुँ अपना बाल बच्चा हमर सब हम अगर सोचबै तऽ हमर बाल बच्चा हमरा जरूर ओ करऽ देतै कि हँ तू माँ कर नीक हेतौ बाल बच्चा सब सब साथ देतै |
| 18 | Manipuri | Not Supported | ꯌꯨ ꯑꯦ ꯑꯦꯟ ꯅꯝꯕꯔ ꯃꯔꯤ ꯃꯔꯤ ꯇꯔꯦꯠ ꯑꯃ ꯃꯥꯄꯜ ꯃꯉꯥ ꯃꯉꯥ ꯇꯔꯦꯠ ꯇꯔꯦꯠ ꯅꯤꯄꯥꯜ ꯃꯥꯄꯜ ꯅꯤꯄꯥꯜ ꯁꯤꯖꯤꯟꯅꯗꯨꯅ ꯑꯩꯒꯤ ꯄ꯭ꯔꯣꯕꯤꯗꯦꯟ ꯐꯟ ꯑꯦꯀꯥꯎꯟꯀꯤ ꯏ ꯄꯤ ꯑꯦꯐ ꯑꯣ ꯄꯥꯁꯕꯨꯛ ꯕꯦꯂꯦꯟꯁ ꯆꯦꯛ ꯇꯧꯕ ꯄꯥꯝꯃꯤ |
| 19 | Nepali | Not Supported | चोकटा खान गएको बुढी जोरमा डुबेर मरि। |
| 20 | Sanskrit | Not Supported | म प्रशासनेन केन्द्रीय कर्मकरेभ्य सेवोत्तरवेतनभोगिभ्यश्च जनेभ्यो महार्घताभृतौ प्रतिशतं पचाशत्मिता वृद्धि समुद्घोषित |
| 21 | Santali | Not Supported | ᱟᱨ ᱥᱮᱨᱢᱟ ᱨᱮᱭᱟᱝ ᱡᱚᱛᱚ ᱠᱷᱚᱱ ᱢᱟᱨᱟᱝ ᱯᱚᱨᱚᱵᱽ ᱫᱚ ᱥᱚᱨᱦᱟᱭ ᱥᱮ ᱵᱟᱦᱟ |
| 22 | Sindhi | Not Supported | ब्लेड जे मयार जी, साराह छो न कई। |
| 23 | Urdu | Not Supported | برسات کے موسم میں بوندوں کا افسوس کرنا  خوبصورتی ہوتا ہے۔ |

Points

**Step 1:**

**Whisper for LID English/Non-English**

1. We are using whisper-small, but it is detecting some of the indic languages as English in the first step itself. Maybe whisper-large can correctly detect, but a large model is not needed for a binary classification. (but will check that)
2. So, we are planning to replace Whisper-small for English/Non-English language identification with SpeechBrain LID/ whisper-medium/large. Will check the results using SpeechBrain. If that doesn't work, will find other options.

**Step 2:**

**LID**

1. Needs to add some fallbacks for LID of Dogri, since Dogri is always detected as Punjabi.
2. Wrong LID performance - Bodo, Dogri (Punjabi), Konkani (Marati), Manipuri (Nepali), Maithili, Kashmiri
3. English is detecting with 100% accuracy (the data contains foreign accent, not indian accent)
4. Will try LID for English with Indian Accent dataset and if that works, will remove step 1.
5. Trying other models for LID of these languages
6. LANGUAGE-WISE ACCURACY FOR MMS-LID 1024:

----------------------------------------------------------------------

Code | Language Name | Files | Top-1 | Top-5 | Conf

----------------------------------------------------------------------

sa | Sanskrit | 41 | 36.6% | 100.0% | 0.670

hi | Hindi | 37 | 86.5% | 94.6% | 0.894

kok | Konkani | 13 | 0.0% | 15.4% | 0.593

sat | Santali | 28 | 25.0% | 39.3% | 0.503

ta | Tamil | 182 | 97.8% | 98.4% | 0.978

pa | Punjabi | 31 | 87.1% | 100.0% | 0.922

en | English | 37 | 100.0% | 100.0% | 0.978

bn | Bengali | 277 | 98.2% | 100.0% | 0.990

kn | Kannada | 40 | 97.5% | 97.5% | 0.970

ur | Urdu | 15 | 86.7% | 100.0% | 0.779

te | Telugu | 119 | 55.5% | 84.0% | 0.763

as | Assamese | 30 | 100.0% | 100.0% | 0.999

ne | Nepali | 31 | 77.4% | 87.1% | 0.869

sd | Sindhi | 180 | 80.6% | 93.9% | 0.822

mai | Maithili | 28 | 0.0% | 0.0% | 0.670

ks | Kashmiri | 35 | 5.7% | 42.9% | 0.258

mni | Manipuri | 25 | 8.0% | 16.0% | 0.296

doi | Dogri | 30 | 0.0% | 0.0% | 0.925

or | Odia | 35 | 65.7% | 65.7% | 0.738

ml | Malayalam | 36 | 72.2% | 88.9% | 0.868

mr | Marathi | 18 | 100.0% | 100.0% | 0.969

gu | Gujarati | 29 | 93.1% | 100.0% | 0.887

br | Bodo | 36 | 5.6% | 36.1% | 0.442

**Total files evaluated: 1333**

**Overall Top-1 Accuracy: 73.89% (985/1333)**

**Overall Top-5 Accuracy: 84.55% (1127/1333)**

2. voxlect-indic-lid-whisper-large-v3

| Language | Accuracy (%) |

| --------- | ------------ |

| Assamese | 86.7 |

| Bengali | 90.0 |

| Bodo | 86.1 |

| Dogri | 96.7 |

| Gujarati | 82.8 |

| English | 100.0 |

| Hindi | 81.0 |

| Kannada | 97.5 |

| Konkani | 30.8 |

| Kashmiri | 20.0 |

| Maithili | 96.4 |

| Malayalam | 69.4 |

| Manipuri | 76.0 |

| Marathi | 44.4 |

| Nepali | 67.7 |

| Odia | 48.6 |

| Punjabi | 9.7 |

| Sanskrit | 97.6 |

| Santali | 85.7 |

| Sindhi | 26.0 |

| Tamil | 83.0 |

| Telugu | 70.0 |

| Urdu | 93.3 |

3. VoxLingua107 ECAPA-TDNN

Total files evaluated: 1333

Overall Top-1 Accuracy: 67.97% (906/1333)

📈 LANGUAGE-WISE ACCURACY:

------------------------------------------------------------

Code | Language Name | Files | Accuracy | Conf

------------------------------------------------------------

sa | Sanskrit | 41 | 12.2% | 0.606

hi | Hindi | 37 | 86.5% | 0.776

kok | Konkani | 13 | 0.0% | 0.718

sat | Santali | 28 | 0.0% | 0.582

ta | Tamil | 182 | 91.8% | 0.937

pa | Punjabi | 31 | 38.7% | 0.605

en | English | 37 | 94.6% | 0.921

bn | Bengali | 277 | 97.1% | 0.969

kn | Kannada | 40 | 100.0% | 0.988

ur | Urdu | 15 | 66.7% | 0.806

te | Telugu | 119 | 64.7% | 0.733

as | Assamese | 30 | 100.0% | 0.996

ne | Nepali | 31 | 77.4% | 0.799

sd | Sindhi | 180 | 78.9% | 0.833

mai | Maithili | 28 | 0.0% | 0.723

ks | Kashmiri | 35 | 0.0% | 0.560

mni | Manipuri | 25 | 0.0% | 0.602

doi | Dogri | 30 | 0.0% | 0.855

or | Odia | 35 | 0.0% | 0.513

ml | Malayalam | 36 | 61.1% | 0.782

mr | Marathi | 18 | 83.3% | 0.812

gu | Gujarati | 29 | 89.7% | 0.899

br | Bodo | 36 | 0.0% | 0.634

1. LANGUAGE-WISE ACCURACY FOR MMS-LID 256:

----------------------------------------------------------------------

Code | Language Name | Files | Top-1 | Top-5 | Conf

----------------------------------------------------------------------

as | Assamese | 30 | 100.0% | 100.0% | 1.000

mr | Marathi | 18 | 100.0% | 100.0% | 0.999

en | English | 37 | 100.0% | 100.0% | 0.980

bn | Bengali | 277 | 99.3% | 100.0% | 0.996

ta | Tamil | 182 | 97.8% | 97.8% | 0.982

kn | Kannada | 40 | 97.5% | 100.0% | 0.993

gu | Gujarati | 29 | 100.0% | 100.0% | 0.978

pa | Punjabi | 31 | 87.1% | 100.0% | 0.941

hi | Hindi | 37 | 73.0% | 97.3% | 0.909

ur | Urdu | 15 | 86.7% | 100.0% | 0.876

sd | Sindhi | 180 | 81.7% | 95.0% | 0.872

ne | Nepali | 31 | 80.6% | 90.3% | 0.874

ml | Malayalam | 36 | 80.6% | 91.7% | 0.892

or | Odia | 35 | 62.9% | 65.7% | 0.753

te | Telugu | 119 | 79.8% | 95.0% | 0.857

sa | Sanskrit | 41 | 80.5% | 97.6% | 0.774

sat | Santali | 28 | 25.0% | 46.4% | 0.658

mni | Manipuri | 25 | 0.0% | 0.0% | 0.545

ks | Kashmiri | 35 | 11.4% | 40.0% | 0.427

br | Bodo | 36 | 0.0% | 0.0% | 0.640

kok | Konkani | 13 | 0.0% | 7.7% | 0.663

mai | Maithili | 28 | 0.0% | 3.6% | 0.799

doi | Dogri | 30 | 0.0% | 0.0% | 0.917

1. LANGUAGE-WISE ACCURACY FOR MMS-LID 512:

LANGUAGE-WISE ACCURACY:

----------------------------------------------------------------------

Code | Language Name | Files | Top-1 | Top-5 | Conf

----------------------------------------------------------------------

sa | Sanskrit | 41 | 85.4% | 100.0% | 0.812

hi | Hindi | 37 | 54.1% | 97.3% | 0.840

kok | Konkani | 13 | 0.0% | 15.4% | 0.499

sat | Santali | 28 | 21.4% | 46.4% | 0.602

ta | Tamil | 182 | 97.8% | 98.4% | 0.982

pa | Punjabi | 31 | 90.3% | 100.0% | 0.950

en | English | 37 | 100.0% | 100.0% | 0.981

bn | Bengali | 277 | 98.2% | 100.0% | 0.992

kn | Kannada | 40 | 97.5% | 97.5% | 0.971

ur | Urdu | 15 | 93.3% | 100.0% | 0.905

te | Telugu | 119 | 58.0% | 88.2% | 0.741

as | Assamese | 30 | 100.0% | 100.0% | 1.000

ne | Nepali | 31 | 80.6% | 90.3% | 0.856

sd | Sindhi | 180 | 66.7% | 87.8% | 0.776

mai | Maithili | 28 | 0.0% | 3.6% | 0.704

ks | Kashmiri | 35 | 8.6% | 31.4% | 0.353

mni | Manipuri | 25 | 8.0% | 28.0% | 0.424

doi | Dogri | 30 | 0.0% | 0.0% | 0.907

or | Odia | 35 | 60.0% | 65.7% | 0.725

ml | Malayalam | 36 | 63.9% | 91.7% | 0.865

mr | Marathi | 18 | 100.0% | 100.0% | 0.997

gu | Gujarati | 29 | 82.8% | 100.0% | 0.874

br | Bodo | 36 | 13.9% | 47.2% | 0.500

| No. | Code | Language | MMS-LID -1024 | MMS-LID -256 | MMS-LID -512 | Voxlect-Whisper  Large-v3 | VoxLingua107 ECAPA-TDNN |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | as | Assamese | 100.0 | 100.0% | 100.0% | 86.7 | 100.0 |
| 2 | mr | Marathi | 100.0 | 100.0% | 100.0% | 44.4 | 83.3 |
| 3 | en | English | 100.0 | 100.0% | 100.0% | 100.0 | 94.6 |
| 4 | bn | Bengali | 98.2 | 99.3% | 98.2% | 90.0 | 97.1 |
| 5 | ta | Tamil | 97.8 | 97.8% | 97.8% | 83.0 | 91.8 |
| 6 | kn | Kannada | 97.5 | 97.5% | 97.5% | 97.5 | 100.0 |
| 7 | gu | Gujarati | 93.1 | 100.0% | 82.8% | 82.8 | 89.7 |
| 8 | pa | Punjabi | 87.1 | 87.1% | 90.3% | 9.7 | 38.7 |
| 9 | hi | Hindi | 86.5 | 73.0% | 54.1% | 81.0 | 86.5 |
| 10 | ur | Urdu | 86.7 | 86.7% | 93.3% | 93.3 | 66.7 |
| 11 | sd | Sindhi | 80.6 | 81.7% | 66.7% | 26.0 | 78.9 |
| 12 | ne | Nepali | 77.4 | 80.6% | 80.6% | 67.7 | 77.4 |
| 13 | ml | Malayalam | 72.2 | 80.6% | 63.9% | 69.4 | 61.1 |
| 14 | or | Odia | 65.7 | 62.9% | 60.0% | 48.6 | 0.0 |
| 15 | te | Telugu | 55.5 | 79.8% | 58.0% | 70.0 | 64.7 |
| 16 | sa | Sanskrit | 36.6 | 80.5% | 85.4% | 97.6 | 12.2 |
| 17 | sat | Santali | 25.0 | 25.0% | 21.4% | 85.7 | 0.0 |
| 18 | mni | Manipuri | 8.0 | 0.0% | 8.0% | 76.0 | 0.0 |
| 19 | ks | Kashmiri | 5.7 | 11.4% | 8.6% | 20.0 | 0.0 |
| 20 | br | Bodo | 5.6 | 0.0% | 13.9% | 86.1 | 0.0 |
| 21 | kok | Konkani | 0.0 | 0.0% | 0.0% | 30.8 | 0.0 |
| 22 | mai | Maithili | 0.0 | 0.0% | 0.0% | 96.4 | 0.0 |
| 23 | doi | Dogri | 0.0 | 0.0% | 0.0% | 96.7 | 0.0 |

| No. | Model | Overall Top-1 Accuracy |
| --- | --- | --- |
| 1 | MMS-LID-1024 | 73.89% |
| 2 | Voxlect-indic-lid-whisper-large-v3 | 71.28% |
| 3 | VoxLingua107 ECAPA-TDNN | 67.97% |
| 4 | MMS-LID-256 | 77.64% |
| 5 | MMS-LID-512 | 72.69% |

Normalised accuracy in terms of statistics based on the number of persons speak ??

1. IndicWhisper
2. CLSRIL-23
3. Voxlect
4. VoxLingua107 wav2vec2 - Aman
5. VoxLingua107 ECAPA TDNN -Jahnavi
6. MMS-LID-256 -Noumida
7. MMS-LID-512-Noumida
8. MMS-LID-1024-Noumida
9. TitaNet-LID (NVIDIA / SpeechBrain)
10. Conformer-LID (TFLite Hub) : indicconformer
11. OWSM CTC v3.1
12. Jiva
13. Wav2Vec2.0-based large XLS-R 3 versions - Ali