Bridging the Gap: Leveraging Cherokee to Improve Language Identification for Endangered Iroquoian Languages

Liam Eggleston, Michael Cacioli, Jatin Sarabu, Kevin Zhu
Algoverse AI Research
kevin@algoverse.us

Abstract

Language identification is a foundational task in natural language processing (NLP), yet many Indigenous languages remain entirely unsupported by commercial language identification systems. In this study, we assess the performance of Google LangID on a 5k Cherokee dataset and find that every sentence is classified as "undetermined", indicating a complete failure to even misidentify Cherokee as another language. To further explore this issue, we manually constructed the first digitalized Northern Iroquoian dataset, consisting of 120 sentences across five related languages: Onondaga, Cayuga, Mohawk, Seneca, and Oneida. Running these sentences through Google LangID, we examine patterns in its incorrect predictions. To address these limitations, we train a random forest classifier to successfully distinguish between these languages, demonstrating its effectiveness in language identification. Our findings underscore the inadequacies of existing commercial language identification models for Indigenous languages and highlight concrete steps toward improving automated recognition of low-resource languages.

1 Introduction

Language identification is fundamental to natural language processing (Kargaran et al., 2023), enabling applications like machine translation, speech recognition, and text classification (Qi et al., 2019). While commercial language technologies such as Google's LangID perform well for high-resource languages, they provide no support for Native American languages (Caswell et al., 2020; Yang et al., 2025b,e). This lack of recognition contributes to digital marginalization and excludes speakers from technological advancements (Bali et al., 2019; Kukulska-Hulme et al., 2023). Cherokee, a Southern Iroquoian language, exemplifies this gap, as it remains computationally underrepresented despite active revitalization efforts

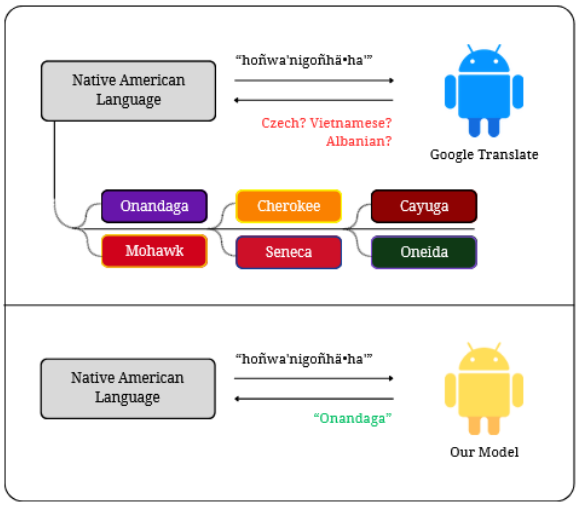

Figure 1: A stylized rendition of our language identification system for endangered Iroquoian languages.

(White, 1962; Peter and Hirata-Edds, 2006; Cushman, 2019).

To investigate this issue, we examined Google LangID's handling of Cherokee and five Northern Iroquoian languages, Onondaga, Cayuga, Mohawk, Seneca, and Oneida, using a manually curated dataset of 120 sentences evenly distributed across languages classes. While Cherokee was consistently misclassified as "undetermined", the other Northern Iroquoian languages were assigned unrelated languages. As shown in Figure 1, we then trained a random forest classifier on Cherokee and these misidentified languages, demonstrating that even with limited data, high classification accuracy is achievable. Our contributions include (1) a novel dataset, (2) an empirical evaluation of Google LangID's misclassification tendencies, and (3) an efficient classification model that outperforms existing approaches.

2 Related Work

Recent NLP research on Indigenous languages has increasingly focused on language identification,

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 1–6, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-001

cross-lingual generalization, and synthetic data generation to mitigate data scarcity. While modern LangID models support hundreds of languages (Kargaran et al., 2023; Milind Agarwal, 2023), they frequently overlook or fail for Indigenous languages due to insufficient training data (Cavalin et al., 2023). One promising approach is family-aware classification, where related languages are incorporated into training. Cavalin et al. (2023) demonstrated this by improving LangID performance for Brazilian Indigenous languages through linguistic family modeling. Similarly, leveraging phonological, morphological, and script-based cues has been proposed as a strategy for improving classification of Cherokee and Northern Iroquoian languages (Kargaran et al., 2023). However, Cherokee's unique syllabary introduces additional challenges compared to the Latin-based scripts used by its linguistic relatives.

Cross-lingual generalization offers a promising approach to improving LangID in low-resource settings. Multilingual models like mBERT can transfer knowledge across related languages (Pires et al., 2019), with pretraining on linguistically similar languages boosting classification accuracy (Bafna et al., 2023). While Cherokee belongs to the Southern Iroquoian branch (Zhang, 2022), it shares structural features with Northern Iroquoian languages, suggesting potential for generalization. However, differences in writing systems may hinder direct transfer, requiring transliteration or character-level modeling (Zhang et al., 2020). Given the scarcity of annotated data, synthetic techniques such as back-translation and morphological augmentation have been explored to enhance NLP models for endangered languages (Feldman and Coto-Solano, 2020; Zhang et al., 2020; Yang et al., 2025c,a). While synthetic data can improve classifier robustness, community validation remains crucial to mitigating risks associated with artificial augmentation (Zhang et al., 2022). Applied thoughtfully, these methods could strengthen language identification for Cherokee and Northern Iroquoian languages.

3 NatAm Language Landscape

The Cherokee language, known as Tsalagi Gawonihisdi (King, 1975), belongs to the Iroquoian language family and is classified under the Southern Iroquoian branch. As shown in Figure 2, it is the only surviving language of this branch (Rountree, 1987), with its closest linguistic relatives found in the Northern Iroquoian group, including Mohawk, Seneca, Oneida, Onondaga, and Cayuga. Linguistic evidence suggests that Cherokee diverged from Northern Iroquoian languages approximately 3,500 to 3,800 years ago (Barrie and Uchihara, 2019), leading to substantial differences in phonology, morphology, and writing systems. Unlike the Northern Iroquoian languages, which primarily rely on oral traditions and Latin-based orthographies (Birch, 2015), Cherokee developed a unique syllabary in the early 19th century (Cushman, 2012), further distinguishing it from its linguistic relatives.

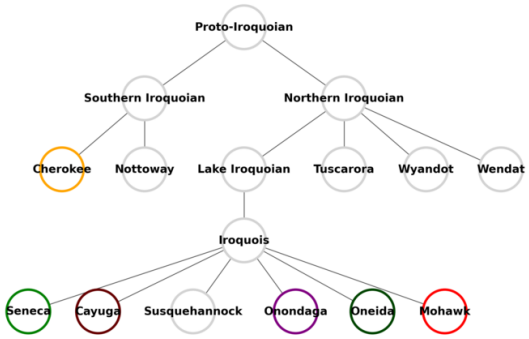

Figure 2: Language family tree for Proto-Iroquoian languages, with Cherokee, Seneca, Cayuga, Onondaga, Oneida and Mohawk highlighted through color.

Mohawk, Seneca, Oneida, Onondaga, and Cayuga, spoken in the northeastern United States and Canada, are members of the Northern Iroquoian branch and share many grammatical and phonological features. Mohawk, one of the most widely spoken Northern Iroquoian languages (Hoover, 1992), has benefited from revitalization programs and digital resources. Seneca and Cayuga, though critically endangered, continue to be taught in community-based initiatives (Chafe, 2015; Dyck and Kumar, 2012). Oneida and Onondaga, while also endangered, have seen growing interest in language preservation efforts through educational programs (Lu et al., 2024; Michelson, 2021). Despite their historical and linguistic connections, these languages exhibit distinct phonetic and syntactic structures (Kilarski, 2021), which may contribute to challenges in language classification. Furthermore, all Iroquoian languages have faced severe endangerment due to colonization and language suppression policies (Richter, 2011), necessitating ongoing revitalization efforts.

4 Data

To assess Google LangID's performance on Cherokee and other Northern Iroquoian languages, we manually collected text samples from publicly available sources[1]. For Cherokee, we were able to refer to an existing 5k dataset (Zhang et al., 2020). Given the scarcity of textual data for the other Northern Iroquoian languages, we manually curated our own digitalized dataset with community-driven language archives, linguistic documentation projects, and publicly available transcripts of Indigenous language programs. Each language was represented by about 20 sentences carefully selected to reflect a range of grammatical structures and vocabulary diversity.

Our decision to rely on manually curated data was driven by the lack of large-scale, digitized corpora for these languages. Automatic web scraping approaches proved ineffective due to the limited online presence of Indigenous languages and difficulty in accurately identifying them, necessitating a more targeted approach to ensure linguistic accuracy and representativeness. Additionally, we prioritized sources produced or validated by native speakers to maintain authenticity and avoid potential biases introduced by machine-generated translations. This novel dataset serves as a foundational resource for evaluating LangID models on Iroquoian languages and underscores the broader challenges of building NLP tools for endangered languages.

5 Language Identification

5.1 Google LangID

To evaluate Google LangID's handling of Cherokee, we passed the 5k Cherokee dataset through the Google Translate API. Surprisingly, every sentence was classified as *undetermined*, meaning the system did not even *attempt* to associate Cherokee with any known language. While Google LangID does not officially support Cherokee, it should at least misidentify it rather than fail to classify it altogether. Prior research on low-resource language identification has shown that unsupported languages are typically misclassified as typologically or phonetically similar ones. For instance, in a recent study on Navajo (Yang et al., 2025d), a 10k dataset was run through Google LangID, and while the results were incorrect, each sentence was still assigned to an existing language. The fact that Cherokee received no such assignment suggests a fundamental failure—not just in recognizing the language, but in engaging with the data at all.

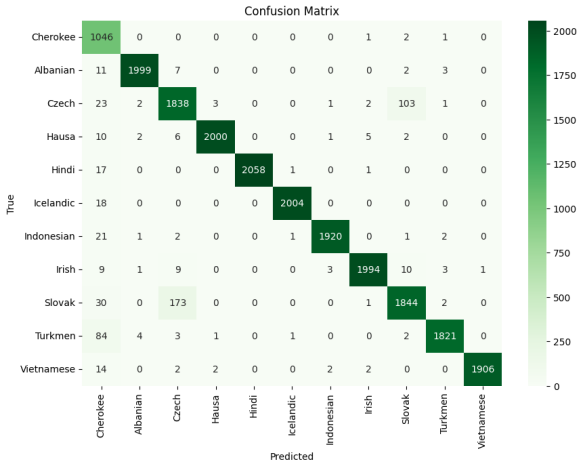

Figure 3: Classification results for Cherokee and 10 other languages, presented as a confusion matrix.

To further investigate, we ran our manually curated dataset of 120 Northern Iroquoian sentences through Google LangID. Unlike Cherokee, these sentences were assigned specific, though incorrect, language labels, indicating that the system at least attempted classification. This stark contrast in performance underscores a deeper issue; while many Indigenous languages are misidentified, Cherokee is uniquely *absent* from the model's processing pipeline, raising concerns about how commercial language technologies handle languages with distinct scripts, such as the Cherokee syllabary.

5.2 Classifier

To address the shortcomings of existing language identification models for Indigenous languages, we developed a custom classifier to distinguish Cherokee from other languages in our dataset. Given the limited availability of labeled data, we selected a Random Forest classifier (Hastie et al., 2009) for its robustness, interpretability, and effectiveness in handling small datasets with high-dimensional features. We employed a TF-IDF vectorizer to transform text into numerical representations, capturing key lexical distinctions. Tokenization was performed at the word level, and the feature space was restricted to the 5,000 most frequent terms to balance specificity and generalization.

The dataset included Cherokee alongside the ten most commonly misidentified languages, such as Albanian, Czech, Hausa, Hindi, Icelandic, Indone-

[1]Full citations are included in the GitHub.

Language	Precision	Recall	F1-Score
Cherokee	0.82	1.00	0.90
Albanian	1.00	0.99	0.99
Czech	0.90	0.93	0.92
Hausa	1.00	0.99	0.99
Hindi	1.00	0.99	1.00
Icelandic	1.00	0.99	0.99
Indonesian	1.00	0.99	0.99
Irish	0.99	0.98	0.99
Slovak	0.94	0.90	0.92
Turkmen	0.99	0.95	0.97
Vietnamese	1.00	0.99	0.99
Accuracy		0.97	
Macro Avg	0.97	0.97	0.97
Weighted Avg	0.97	0.97	0.97

Table 1: Multi-classifier Performance Metrics.

sian, Irish, Slovak, Turkmen, and Vietnamese. Text samples were manually curated and preprocessed to remove extraneous whitespace before vectorization. A stratified 80-20 train-test split ensured balanced representation across all classes. Training was conducted with 100 decision trees, using a fixed random state for reproducibility. Evaluation metrics (precision, recall, and F1-score) demonstrated strong differentiation between Cherokee and the other languages, though minor misclassifications occurred, particularly among typologically similar languages. The confusion matrix in Figure 3 highlights these cases, emphasizing the challenge of distinguishing languages with overlapping linguistic structures. The effectiveness of TF-IDF features in capturing distinguishing characteristics while filtering out noise from infrequent words is further reflected in Table 1.

Further analysis of the model's binary classification performance in Table 2 shows high accuracy in distinguishing Cherokee from all other languages. The precision and recall scores confirm the classifier's reliability in identifying Cherokee while correctly classifying non-Cherokee languages. Our results demonstrate that even with limited training data, a random forest classifier can effectively differentiate Indigenous from non-Indigenous languages, addressing gaps in commercial language identification. Future work could expand the dataset through community-driven contributions, incorporate additional Indigenous languages, and refine feature selection to enhance classification. Exploring deep learning approaches may further improve performance, fostering the development of more inclusive NLP tools for endangered languages.

Class	Precision	Recall	F1-Score
Cherokee	1.00	0.82	0.90
Non-Cherokee	0.99	1.00	1.00
Accuracy		0.99	
Macro Avg	1.00	0.91	0.95

Table 2: Binary Classification Performance Metrics.

6 Future Work

Future research will include interviews with Indigenous community members to gain cultural insights into language classification challenges. We have already scheduled two interviews with an Omaha Tribe member and a member of the Okanagan/Wenatchi community, ensuring direct engagement with native speakers. Expanding the dataset to incorporate additional Indigenous languages and exploring deep learning models will further improve classification accuracy (Alvarez et al., 2025). Additionally, integrating phonetic and morphological features will enhance model interpretability, while ethical considerations will guide meaningful collaboration with Indigenous communities for validation and tool development. These efforts aim to create more inclusive and effective language identification tools that actively support Indigenous language preservation.

7 Conclusion

This study highlights the severe shortcomings of commercial language identification systems for Indigenous languages, exemplified by Google LangID's failure to classify Cherokee—even incorrectly. While other Northern Iroquoian languages received misidentifications, Cherokee was uniquely ignored, raising concerns about how commercial models handle languages with distinct scripts. To address this gap, we developed a random forest classifier that effectively differentiates Cherokee, demonstrating that even with limited data, accurate classification is achievable. Our findings underscore the need for more inclusive NLP tools that support endangered languages. **We call upon the NLP community to move beyond discussion and take concrete steps, whether by expanding datasets or collaborating with Indigenous speakers, to ensure that these languages are not just studied, but actively supported.**

Limitations

While our study provides valuable insights into the deficiencies of commercial LangID models for Cherokee and Northern Iroquoian languages, it is constrained by the small dataset size and the absence of native speaker validation. Additionally, our classifier's effectiveness may not extend to other underrepresented Indigenous languages with different linguistic structures. Further research should explore larger datasets, multimodal approaches, and direct collaboration with Indigenous speakers to improve the accuracy and ethical implementation of language identification systems.

Ethics Statement

Our study prioritizes ethical data collection and representation of Indigenous languages. We sourced data only from publicly available and community-approved resources, ensuring that no proprietary or culturally sensitive materials were used without consent. Additionally, we acknowledge the historical and ongoing marginalization of Indigenous languages in NLP and aim to contribute to language preservation rather than commodification. Future work should actively involve Indigenous communities in data collection and validation to ensure their agency in technological advancements. In the spirit of transparent and ethical research, samples of our data and code has been made available at (https://github.com/Cherokee-Project/Classifier).

References

Jesus Alvarez, Daua Karajeanes, Ashley Prado, John Ruttan, Ivory Yang, Sean O'Brien, Vasu Sharma, and Kevin Zhu. 2025. Advancing uto-aztecan language technologies: A case study on the endangered comanche language. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 27–37.

Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, and Rachel Bawden. 2023. Cross-lingual strategies for low-resource language modeling: A study on five indic dialects. In *18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 28–42. ATALA.

Kalika Bali, Monojit Choudhury, Sunaya Sitaram, and Vivek Seshadri. 2019. Ellora: Enabling low resource languages with technology. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 160–163.

Michael Barrie and Hiroto Uchihara. 2019. Iroquoian languages. In *The Routledge handbook of North American languages*, pages 424–451. Routledge.

Jennifer Birch. 2015. Current research on the historical development of northern iroquoian societies. *Journal of Archaeological Research*, 23:263–323.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.

Paulo Cavalin, Pedro Domingues, Julio Nogima, and Claudio Pinhanez. 2023. Understanding native language identification for brazilian indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 12–18.

Wallace L Chafe. 2015. *A grammar of the Seneca language*, volume 149. Univ of California Press.

Ellen Cushman. 2012. *The Cherokee syllabary: Writing the people's perseverance*, volume 56. University of Oklahoma Press.

Ellen Cushman. 2019. Language perseverance and translation of cherokee documents. *College English*, 82(1):115–134.

Carrie Dyck and Ranjeet Kumar. 2012. A grammar-driven bilingual digital dictionary for cayuga (iroquoian). *Dictionaries: Journal of the Dictionary Society of North America*, 33(1):179–204.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, pages 587–604.

Michael L Hoover. 1992. The revival of the mohawk language in kahnawake. *Canadian Journal of Native Studies*, 12(2):269–287.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.

Marcin Kilarski. 2021. Sound systems in iroquoian languages. In *A History of the Study of the Indigenous Languages of North America*, pages 131–172. John Benjamins Publishing Company.

Duane Harold King. 1975. *A grammar and dictionary of the Cherokee language.* University of Georgia.

Agnes Kukulska-Hulme, Ram Ashish Giri, Saraswati Dawadi, Kamal Raj Devkota, and Mark Gaved. 2023. Languages and technologies in education at school and outside of school: Perspectives from young people in low-resource countries in africa and asia. *Frontiers in Communication*, 8:1081155.

Yanfei Lu, Patrick Littell, and Keren Rice. 2024. Empowering oneida language revitalization: Development of an oneida verb conjugator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5757–5767.

Karin Michelson. 2021. A reference grammar of the onondaga language.

Antonios Anastasopoulos Milind Agarwal, Md. Mahfuz Ibn Alam. 2023. Limit: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *EMNLP 2023*.

Lizette Peter and Tracy E Hirata-Edds. 2006. Using assessment to inform instruction in cherokee language revitalisation. *International Journal of Bilingual Education and Bilingualism*, 9(5):643–658.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Zhaodi Qi, Yong Ma, and Mingliang Gu. 2019. A study on low-resource language identification. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1897–1902. IEEE.

Daniel K Richter. 2011. *The ordeal of the longhouse: the peoples of the Iroquois League in the era of European colonization*. UNC Press Books.

Helen C Rountree. 1987. The termination and dispersal of the nottoway indians of virginia. *The Virginia Magazine of History and Biography*, 95(2):193–214.

John K White. 1962. On the revival of printing in the cherokee language. *Current Anthropology*, 3(5):511–514.

Ivory Yang, Xiaobo Guo, Yuxin Wang, Hefan Zhang, Yaning Jia, William Dinauer, and Soroush Vosoughi. 2025a. Recontextualizing revitalization: A mixed media approach to reviving the nüshu language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Ivory Yang, Weicheng Ma, Carlos Guerrero Alvarez, William Dinauer, and Soroush Vosoughi. 2025b. What is it? towards a generalizable native american language identification system. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 105–111.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025c. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025d. Is it Navajo? accurate language detection for endangered athabaskan languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 277–284.

Ivory Yang, Chunhui Zhang, Yuxin Wang, Zhongyu Ouyang, and Soroush Vosoughi. 2025e. Visibility as survival: Generalizing nlp for native alaskan language identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6965–6979.

Bryan Zhang. 2022. Improve MT for search with selected translation memory using search signals. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 123–131, Orlando, USA. Association for Machine Translation in the Americas.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595.