World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:19, No:01, 2025

# Detecting Indigenous Languages: A System for Maya Text Profiling and Machine Learning Classification Techniques

**Authors :** Alejandro Molina-Villegas, Silvia Fernández-Sabido, Eduardo Mendoza-Vargas, Fátima Miranda-Pestaña

**Abstract :** The automatic detection of indigenous languages in digital texts is essential to promote their inclusion in digital media. Underrepresented languages, such as Maya, are often excluded from language detection tools like Google's language-detection library, LANGDETECT. This study addresses these limitations by developing a hybrid language detection solution that accurately distinguishes Maya (YUA) from Spanish (ES). Two strategies are employed: the first focuses on creating a profile for the Maya language within the LANGDETECT library, while the second involves training a Naive Bayes classification model with two categories, YUA and ES. The process includes comprehensive data preprocessing steps, such as cleaning, normalization, tokenization, and n-gram counting, applied to text samples collected from various sources, including articles from La Jornada Maya, a major newspaper in Mexico and the only media outlet that includes a Maya section. After the training phase, a portion of the data is used to create the YUA profile within LANGDETECT, which achieves an accuracy rate above 95% in identifying the Maya language during testing. Additionally, the Naive Bayes classifier, trained and tested on the same database, achieves an accuracy close to 98% in distinguishing between Maya and Spanish, with further validation through F1 score, recall, and logarithmic scoring, without signs of overfitting. This strategy, which combines the LANGDETECT profile with a Naive Bayes model, highlights an adaptable framework that can be extended to other underrepresented languages in future research. This fills a gap in Natural Language Processing and supports the preservation and revitalization of these languages.

**Keywords :** indigenous languages, language detection, Maya language, Naive Bayes classifier, natural language processing, low-resource languages
**Conference Title :** ICICS 2025 : International Conference on Information and Computer Sciences
**Conference Location :** Cancun, Mexico
**Conference Dates :** January 23-24, 2025