

Nahuatl Dialectal Varieties Classification^{*}

Juan-José Guzmán-Landa^{1,†}, Daniela Juárez-Morales^{2,†}, Juan-Manuel Torres-Moreno^{1,†},
Luis-Gil Moreno-Jiménez^{3,*,†}, Miguel Figueroa-Saavedra^{2,†}, Rémy Kessler^{1,†},
Graham Ranger^{1,†} and Martha-Lorena Avendaño-Garrido^{2,†}

¹LLA/ICTT/Agorantic Université d'Avignon (France)

²Universidad Veracruzana (Mexico)

³INRIA Grenoble (France)

Abstract

Nahuatl, Mexico's indigenous national language, is spoken by more than 2 million people. Despite its rich cultural heritage dating back to the 15th century, Nahuatl is a language with very few computer resources. The problem is even greater because there are approximately 30 recognized dialectal varieties, not counting the different spellings in the language's written forms. In this study, we address an NLP problem concerning fine-grained classification in Nahuatl dialectal varieties, using classical and efficient methods such as TextCat, SVM and Decision Trees; Neural algorithms such as LSTM, C-LSTM, Transformer encoder and xlmRoBERTa. The results obtained show excellent performance (Dialectal varieties F1-score=91.7%) in fine-grained classification tasks.

Keywords

Nahuatl, Dialectal varieties, Classification, Low-resourced Languages, Transformers

1. Introduction

Languages with few digital resources constitute a large proportion of the languages spoken in the world [1] and are distributed across large geographical regions. However, for a variety of reasons (political, economic, or social), they have been technologically neglected in Natural Language Processing (NLP). Consequently, these languages or π -languages have low-or extremely low-computational resources and tools for automatic processing. For example, the Nahuatl or Nahuatl language has existed for thousands of years. It is one of Mexico's national indigenous languages, with the second-highest number of speakers, at approximately 1.65 million people¹; however, Nahuatl is a very low-resource language. The language has considerable dialectal diversity categorised into four main geographical groups²: Western, Central, Eastern, and Huasteca [2]. The Mexico's INALI³ recognises 30 varieties of Nahuatl distributed across different regions of Mexico. These groups are established according to phonological, lexical, and semantic features that reflect their variability, different processes of change, differentiation, and convergence, which are not unrelated to contact between varieties and to processes of specialization, intellectualism, and social prestige. This linguistic diversity poses challenges for the use of corpora in educational, communicative, and digital contexts, as it involves significant variation in spelling and lexical selection [3]. The process is all the more obvious in contexts of written communication where different literacy proposals have also been produced over time without a full standardization process taking place. This adds another characteristic to the written record of oral texts - and therefore their circulation on the Internet - which is the diversity of alphabetic writing systems used. In relation to the dialect variety, this does not necessarily mean that each variety has its own

Posters, Demos, Blue Sky, and Tutorials at SEMANTICS 2026, Sep 2026, Ghent, Belgium

^{*}Corresponding author.

✉ juan-jose.guzman-landa@univ-avignon.fr (J. Guzmán-Landa); danielajum02@gmail.com (D. Juárez-Morales);
juan-manuel.torres@univ-avignon.fr (J. Torres-Moreno); lmoreno30470@gmail.com (L. Moreno-Jiménez);
migfigueroa@uv.mx (M. Figueroa-Saavedra); remy.kessler@univ-avignon.fr (R. Kessler); graham.ranger@univ-avignon.fr
(G. Ranger); maravendano@uv.mx (M. Avendaño-Garrido)

ORCID 0009-0003-7016-9629 (J. Guzmán-Landa); 0000-0002-4392-1825 (J. Torres-Moreno); 0000-0001-7753-7349
(L. Moreno-Jiménez); 0000-0001-5990-1258 (M. Figueroa-Saavedra); 0000-0002-9947-3048 (R. Kessler)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Censo de población y vivienda 2020, INEGI: <https://www.inegi.org.mx>

²<https://www.ethnologue.com>

³https://www.inali.gob.mx/clin-inali/html/l_nahuatl.html

writing, since the same variety can be written with several alphabets and two varieties can share an alphabet. Even throughout history speakers and writers of a variety may have changed systems. In the hope that new social and political conditions surrounding language policies will enable greater use, acceptance, and appreciation of mutual differences, the emergence of a multiplicity of cultural products in different varieties requires resources to facilitate communication and enable these products to reach a larger consumer base. Thus, an NLP tool that could identify and classify different document types based on their linguistic and lexical characteristics would be a valuable resource to support other tools, such as translators, emotion detectors, etc. These tools may empower and strengthen the community of users of this language by making information more accessible without requiring users to give up their mother tongue to access other varieties. Our objectives are to adapt and combine various multi-class classifiers to identify different dialectal varieties of written Nawatl with high precision.

2. Dialectal varieties of Nawatl

Nawatl is an agglutinative and polysynthetic language, a feature that is formally reflected in the use of long words that convey multiple nuances of meaning [4]. It can even give rise to veritable sentence-words, where a verbal word integrates both the subject and the predicate of a typical sentence in many Indo-European languages. This π -language belongs to the Yutonahuas language group. Initially located in the southwestern U.S., Nawatl-speaking groups gradually moved southward in successive waves, settling in territories now part of Mexico and Central America (Guatemala, El Salvador, and Nicaragua). Thus, since the 5th century, Nawatl linguistic groups such as Pochuteco (now extinct) and Nawatl have emerged, with the latter dividing into the western-central Nawatl group and the eastern-northern Nawatl group [5]. Each of these groups came into contact with different environments and other linguistic and cultural groups. This historical development is reflected in the great morphosyntactic and lexical-semantic variability that characterizes speech at regional and biocultural levels. Given this reality, an automatic variety classification tool is becoming increasingly relevant. In Nawatl, there are varieties with more differences, for example, some variations are morphological, relating to how words are constructed, whether nouns or verbs, in terms of agglutination. Thus, we can see an example in the conjugation of imperfective and perfective tenses, where different variants opt for different ways of forming them, both in the selection of the verb stem and the verbal suffixes to incorporate. Thus, there is variability in the use of verbal suffixes that mark the perfective tense in the singular (-Ø, -k, -ki) and the imperfective tense (-ya, -yaya, -tuya), and the incorporation of an aorist augment (o) that places both the perfective and imperfective forms closer to the past tense, while other variants consider the imperfective tense as closer to the present tense. A well-established example is the use of certain verbs. Thus, the verb “to be” can be expressed as *kah* or *yetok* in the Western and Central varieties, as *unkah* or *onkah* in the Nawatl of Guerrero, whereas in the Eastern and South *ono* is used, and in the Huasteca region, two forms are used: *eltok* for inanimate subjects and *itstok* for animate subjects. In this way, the phrase “I am” can be rendered as *nikah*, *nunkah*, *nono*, *niitstok*. At the phraseological level, the varieties show real diversity and complexity (Table 1). Thus, whether they display variations in form, construction, or meaning, it is critical to identify which varieties are involved. Hence, there is a need for an efficient classifier of dialectal varieties.

Table 1
Phraseological examples

Regions	There was a married man who had a wife
Western	<i>Niman kataya se lakal munamiktijtuk kipiataya isiua</i>
Central	<i>Melak yokatka sentetl tlakatl monamiktitok kipiaya un isowah / Nelli katki se tlakatl monamiktihok okipiyaya in isiwaw</i>
Huasteca	<i>Nelia itstoya se tlakatl kipiayaya ni isiwaj</i>
Eastern	<i>Nel onoya se tagat kipiaya monamiktitok ipalmiya / Nel nemik se takat munamiktijtuk kipiatusa ne isiwaw</i>

3. Nawatl corpora

Currently, there are few Nawatl corpora that are truly exploitable for NLP tasks. An important achievement is the corpus of classical Nawatl, developed at UNAM and integrated into the “Gran Diccionario Náhuatl”⁴. On other hand, the “Universal Dependencies Treebank for Highland Puebla Nawatl” corpus [6], focuses on the Sierra de Puebla Nawatl, with an extremely limited number of sentences and several mistakes. It is clear that factors such as the oral nature of Nawatl, the lack of standardisation in spelling, and the existence of around 30 varieties mean that the number of available resources is not particularly high. Among the different corpora, the π -yalli corpus has interesting features in the present context⁵. This is a heterogeneous –but small– Nawatl corpus that is suitable for training of static embeddings. However, this corpus has never been fully characterized, and no attempt has been made to classify its different dialectal varieties. We have decided to manually characterise the π -yalli corpus in order to use it in our experiments. Hereafter, we use the following acronyms as labels for the 25 dialectal varieties available in the π -yalli corpus⁶:

CV	Central de Veracruz	GUE	Guerrero	SNP	Sierra Negra de Puebla
H	Huasteca (a heterogeneous mixture)	NOC	Noroeste Central	SNNP	Sierra Negra Norte Puebla
HH	Huasteca Hidalguense	IST	Istmo	SNSP	Sierra Negra Sur de Puebla
HV	Huasteca Veracruzana	NAW	Nawatl	SNEP	Sierra Noreste de Puebla
HP	Huasteca Potosina	OAX	Oaxaca	SOP	Sierra Oeste de Puebla
CEA	Centro Alto	CEP	Centro de Puebla	TEM	Temixco
CEO	Centro de Occidente	ANP	Alto del Norte de Puebla	TLA	Tlaxcala
CEN	Centro	ORP	Oriente de Puebla	CEB	Central Bajo

The π -yalli corpus contains several million tokens⁷. Often, each token represents a “word-sentence” in Nawatl. The basic statistics of the corpus are as follows: 4 746 documents; $\approx 6\,629\,000$ tokens; $\approx 428\,000$ sentences and 25 dialectal varieties⁸. Figure 1 (left) shows the number of tokens on the $k=18$ dialectal varieties from the π -yalli corpus. The varieties **HV** “Huasteca Veracruzana”, **HP** “Huasteca Potosina” and **CEA** “Centro Alto” dominate the top three places. These first 18 varieties represent, in terms of the number of tokens, 99.7% of corpus.

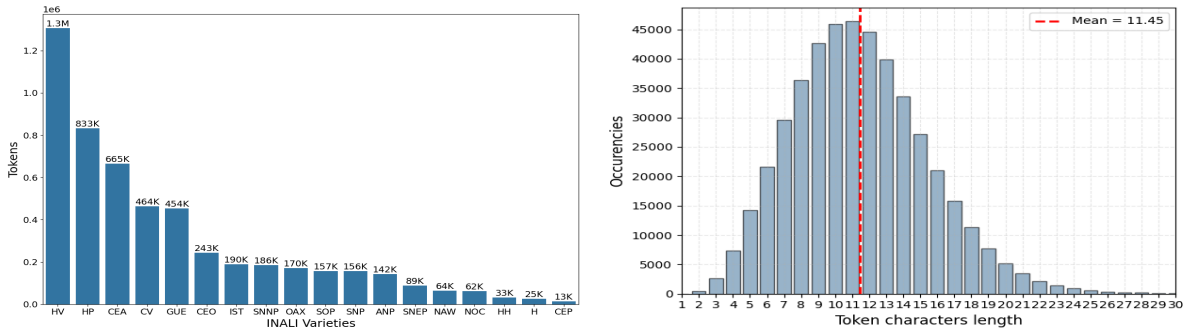


Figure 1: At left: # of tokens concerning the 18 most representative varieties Nawatl from π -yalli corpus. At right: average length of the tokens with respect to the vocabulary of π -yalli corpus

4. Related works

The classification of dialectal varieties can be tackled as a multi-class classification problem. Some studies addressing classification tasks include models designed to identify dialects in languages such as Spanish, English, or Portuguese [7]. Other works have studied the variations among 10 varieties of Nawatl based on: mutual intelligibility, genetic similarity and written varieties [8]. A character language model using a LSTM algorithm was trained with publicly available Bible translations⁹. The

⁴Universidad Nacional Autónoma de México, <https://gdn.iib.unam.mx/>

⁵<https://demo-lia.univ-avignon.fr/pi-yalli>

⁶We used the INALI classification: <https://atlas.inali.gob.mx/agrupaciones/info/0211>

⁷Sequences of characters separated by blank spaces.

⁸ π -yalli is divided into 16 thematic varieties, but this classification is beyond the scope of the current paper.

⁹The sentences are short verses of religious texts from <https://scriptureearth.org>

findings indicate 99% similarity between groups of dialectal varieties under optimal conditions, on only 3 Nawatl varieties using mBERT [9]. The analysis of linguistic varieties in languages with massive computational resources can be tackled directly using Transformer-type models (BERT, mBERT, etc.) [7]. However, other algorithms can also be used, such as LSTM [10] and Siamese neural networks [11]. Even models with simpler architectures offer competitive performance: convolutional neural networks (CNNs) [12], approaches based on character n -grams [13]; SVM classifiers [14] and Decision trees. All of these latter methods are generally better suited to working with small corpora.

5. Experimental setup, dataset and algorithms

All documents in the π -yalli corpus are labelled according to their dialectal variety. We process only the $k=18$ dialectal varieties $> 10K$ tokens (Figure 1, left). The texts were split into “word-sentences” using a segmentation heuristic based on punctuation, proper names, and certain exceptions. Subsequently, there are $\approx 374K$ sentences. To mitigate potential bias on training, we performed cross-validation (80% training; 20% validation, 5 iterations) using 90% corpus. On the test we used the remaining 10% (37.4K sentences) of the corpus as test dataset, in order to evaluate each algorithm individually. We employed 2 texts representations: character n -grams and word embeddings. Character n -grams capture the lexical diversity in a nuanced way. Word Embeddings produces vector representations that capture contextual information in reduced dimensions. FastText embeddings (CBOW or Skip-gram) have shown to be effective in semantic similarity tasks in Nawatl [15], and we decided to use them as text representation for classifiers. **Baseline:** TextCat is a classic yet robust algorithm using character n -grams to identify languages. It is an efficient algorithm and has low computational costs because it uses few resources and computations [13]. **LightGBM:** it is a popular decision tree-based algorithm due to its efficiency in terms of execution time and its good performance¹⁰. **SVM:** are highly efficient multi-class classification algorithms for high-dimensional vectors [14]. The main idea is to find the optimal margin separating hyperplane. **CNN:** are algorithms primarily focused on image processing [16]. In effect, convolution seeks to capture and regroup existing information in image segments. However, textual information can also be processed in a similar manner [17, 18]. From textual segments of m words, information can be extracted by applying convolution. **LSTM:** is a recurrent Neural Network with dedicated memories [10]. The model learns to retain important information despite having processed a large number of textual patterns [19]. Siamese Networks (**C-LSTM**) are architectures that may combine LSTM recurrent networks and CNN networks or others [11]. Siamese networks have been successfully applied to textual similarity tasks. **TREN:** It is the encoder part of a transformer. We focused on their attention mechanism and its capacity to capture diverse semantic information. This entire architecture [20] is well known for its ability to process all tokens simultaneously rather than sequentially. They do not lose their positional information thanks to positional encoding. **xlm-RoBERTa:** was pre-trained in over 100 languages. In our experiments, we fine-tuned xlm-RoBERTa¹¹ using the π -yalli corpus. We employed the standard hyperparameters used in the literature [21, 22]. As a final strategy, the outputs of algorithms were combined into a decision algorithm via a **Soft vote**. Calculates the average vector of all probabilities and uses the maximum value.

6. Results

The baseline achieves a low F1-score $\approx 76\%$. The lowest-ranked class is **H**, with F1-Score $\approx 24\%$. **lightGBM** achieves a F1-Score $\approx 84\%$, with the advantages of producing interpretable rules and being the fastest algorithm. **SVM_R** algorithm achieves good average F-Score $\approx 89\%$. The neural models **CNN**, **LSTM** and **C-LSTM** achieve F1-scores of between 87.4% and 89.6%. **TREN** and **xlmRoBERTa** network achieve similar good performances: (Precision, Recall, F1-Score) $\approx (92\%, 89\%, 90\%)$. However using **Soft Vote**, significantly improved the results of the individual algorithms (average F-Score = 91.7%), see Table 2.

¹⁰<https://lightgbm.readthedocs.io/en/latest/Python-Intro.html>

¹¹Base version with 270 million parameters from: https://huggingface.co/docs/transformers/model_doc/xlm-roberta

Table 2

F1-score 80/20 Merged cross-validation of 90% of the labelled samples. The average results on the Test set (the remaining 10%). VOTE = Soft voting method.

	TextCat	lightGBM	SVM _L	SVM _R	CNN	LSTM	C-LSTM	TREN	RoBERTa	VOTE
Variety	F1-Score									
HV	0.804	0.928	0.932	0.753	0.943	0.954	0.946	0.956	0.963	0.959
HP	0.845	0.930	0.923	0.748	0.952	0.958	0.951	0.955	0.964	0.962
GUE	0.869	0.961	0.963	0.976	0.967	0.975	0.968	0.966	0.972	0.981
CEA	0.886	0.940	0.948	0.962	0.953	0.962	0.955	0.959	0.962	0.968
CV	0.768	0.898	0.899	0.918	0.914	0.926	0.917	0.923	0.920	0.934
CEO	0.984	0.993	0.992	0.995	0.993	0.995	0.994	0.994	0.994	0.996
SNNP	0.867	0.914	0.934	0.952	0.941	0.954	0.943	0.952	0.948	0.963
ANP	0.950	0.970	0.981	0.985	0.981	0.988	0.984	0.988	0.986	0.991
OAX	0.956	0.955	0.965	0.982	0.974	0.982	0.972	0.981	0.976	0.984
SNP	0.938	0.943	0.961	0.974	0.965	0.973	0.965	0.975	0.972	0.981
SOP	0.932	0.969	0.973	0.986	0.977	0.981	0.979	0.982	0.977	0.986
IST	0.968	0.982	0.987	0.994	0.990	0.988	0.989	0.984	0.992	0.996
NAW	0.964	0.953	0.951	0.960	0.953	0.959	0.957	0.963	0.966	0.968
SNEP	0.587	0.681	0.728	0.793	0.713	0.764	0.738	0.777	0.752	0.800
NOC	0.531	0.734	0.775	0.850	0.768	0.817	0.805	0.808	0.789	0.840
H	0.235	0.408	0.401	0.686	0.379	0.521	0.402	0.620	0.618	0.682
HH	0.243	0.509	0.419	0.696	0.669	0.726	0.676	0.713	0.717	0.736
CEP	0.386	0.404	0.641	0.765	0.704	0.709	0.687	0.742	0.777	0.776
Mean	0.762	0.837	0.854	0.887	0.874	0.896	0.879	0.902	0.902	0.917

7. Conclusions and Future Work

In this article, we show that the F1-Score obtained using soft voting outperformed the F1-Score obtained using any model individually. Our hypothesis is that soft voting utilises all the probabilities of each prediction, rather than just the highest probability. In other hand, xlmRoBERTa utilises knowledge acquired across 100 languages and is then fine-tuned using π -yalli corpus, whilst the other models are trained directly using only π -yalli. This enables it to achieve competitive performance; however, more economical models can match and even surpass its performance.

The dialectal classification of Nahuatl can be approached as a classic multi-class NLP problem. We show that, although it is not a straightforward task, the combination of representations and different learning algorithms yields very good results. In fact, amongst the 18 dialectal varieties best represented in the corpus studied, the baseline method based on n -gram character features does obtain poor results, being far outperformed by the SVM and CNN methods. The LSTM network offers the best results, achieving an F1-score of 89.6%. Meanwhile, TREN and RoBERTa achieve the highest average F1-Score: 90.2%. However, when the results are combined using soft voting, the individual results are surpassed, yielding an F1-Score of 91.7%. This illustrates the importance of combining different algorithmic approaches to tackle a NLP problem. In this way, not only is the feasibility of this study’s practical objective confirmed, but also its applicability to the adaptation and translation processes of Nahuatl-speaking communities that are in contact with the rest of the world. As future work, we will explore the possibility of incorporating additional classification systems, and use **Knowledge Graphs & Ontologies** based on Wikidata/Wikipedia Nawatl in order to extend our experiments to the 7 currently under-represented varieties, and also to classify the different topics presents in the corpus.

Acknowledgments

This work was financed by Agorantic/LIA/Avignon Université (France).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] V. Berment, Méthodes pour informatiser les langues et les groupes de langues “peu dotées”, Ph.D. thesis, Université Joseph-Fourier - Grenoble I, 2004.
- [2] U. Canger, Nahuatl dialectology, *Int. Journal of American Linguistics* 54 (1988) 28 – 72.
- [3] M. P. Hansen, *Nahuatl Nations: Language Revitalization and Semiotic Sovereignty in Indigenous Mexico*, Oxford University Press, 2024.
- [4] D. De Felice, Nahuatl as a classical, foreign, and additional language: A phenomenological study, *The Qualitative Report* 17 (2012) 1–23.
- [5] L. Campbell, R. Langacker, Proto-aztecán vowels: Part i, *I. J. of Amer. Linguistics* 44 (1978) 85–102.
- [6] R. Pugh, F. Tyers, A Universal Dependencies treebank for Highland Puebla Nahuatl, in: *NAACL: Human Language Technologies (Vol 1)*, ACL, Mexico, 2024, pp. 1393–1403.
- [7] A. Vaidya, A. Kane, Two-stage pipeline for multilingual dialect detection, in: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, ACL, Dubrovnik, 2023, pp. 222–229.
- [8] R. Pugh, F. Tyers, Investigating variation in written forms of Nahuatl using character-based language models, in: *Workshop on NLP for Indigenous Languages of the Americas*, ACL, 2021, pp. 21–27.
- [9] R. Pugh, F. Tyers, Experiments in multi-variant natural language processing for Nahuatl, in: *11th Workshop on NLP for Similar Languages, Varieties, and Dialects*, ACL, Mexico, 2024, pp. 140–151.
- [10] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (1997) 1735–1780.
- [11] C. Zhou, et al, A C-LSTM neural network for text classification, *CoRR* abs/1511.08630 (2015).
- [12] J. Mairal, et al, Convolutional kernel networks, in: *NIPS*, MIT Press, Cambridge, 2014, p. 2627–2635.
- [13] K. Hornik, et al, The textcat package for *n*-gram based text categorization in R, *J. of Statistical Software* 52 (2013) 1–17.
- [14] C. Cortes, V. Vapnik, Support-vector networks, *Machine learning* 20 (1995) 273–297.
- [15] J.-J. Guzmán-Landa, et al, π -YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl, in: *TALN*, v1, Marseille, 2025, pp. 802–816.
- [16] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), *EMNLP*, ACL, Doha, Qatar, 2014, pp. 1746–1751.
- [17] R. Johnson, T. Zhang, Deep pyramid convolutional neural networks for text categorization, in: *55th Annual Meeting of the ACL (Vol 1: Long Papers)*, ACL, Vancouver, Canada, 2017, pp. 562–570. doi:10.18653/v1/P17-1052.
- [18] A. Jacovi, et al, Understanding convolutional neural networks for text classification, in: *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting NN for NLP*, ACL, Brussels, 2018, pp. 56–65. doi:10.18653/v1/W18-5408.
- [19] S. Jamshidi, et al, Effective text classification using bert, mtm lstm, and dt, *DKE* 151 (2024).
- [20] A. Vaswani, et al, Attention is all you need, 2023. arXiv:1706.03762.
- [21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *NACAAACL: HLT*, Vol. 1, Minneapolis, 2019, pp. 4171–4186.
- [22] A. Conneau, et al, Unsupervised cross-lingual representation learning at scale, *CoRR* abs/1911.02116 (2019). URL: <http://arxiv.org/abs/1911.02116>. arXiv:1911.02116.