Analysis of a Brazilian Indigenous corpus using machine learning methods

Tiago Barbosa de Lima[1]0000-0002-0707-522X , André C. A. Nascimento[1] , Pericles Miranda[1] , Rafael Ferreira Mello[1]

[1]Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, Recife, Pernambuco, 52171-900, Brazil

{tiago.blima,andre.camara,rafael.mello}@ufrpe.br,periclesmiranda@gmail.com

Abstract. *In Brazil, several minority languages suffer a serious risk of extinction. The appropriate documentation of such languages is a fundamental step to avoid that. However, for some of those languages, only a small amount of text corpora is digitally accessible. Meanwhile there are many issues related to the identification of indigenous languages, which may help to identify key similarities among them, as well as to connect related languages and dialects. Therefore, this paper proposes to study and automatically classify 26 neglected Brazilian native languages, considering a small amount of training data, under a supervised and unsupervised setting. Our findings indicate that the use of machine learning models to the analysis of Brazilian Indigenous corpora is very promising, and we hope this work encourage more research on this topic in the next years.*

1. Introduction

The language documentation process of Brazilian Indigenous languages is recent [Moore and Galucio 2016]. Depending on the classification criteria adopted, there are between 160 to 180 Brazilian Indigenous languages in Brazil [Drude et al. 2007, Moore and Galucio 2016]. Most of them suffer the risk of extinction by the end of this century [Drude et al. 2007], making it relevant to encourage more research on the documentation of such languages. Despite the fact that Language Classification (LC) is well established for languages spoken by a larger part of the population, indigenous languages have not received much attention, with very few studies focused in such idiomatic groups [Drude et al. 2007, Moore et al. 2008, Jauhiainen et al. 2019b]. Therefore, automatic language identification might improve the process of categorising those languages since more data could be collected in a short amount of time.

Artificial Intelligence (AI) algorithms have already been proven to be able to accurately categorise a diverse number of languages [Jauhiainen et al. 2019b]. Nonetheless, it is not always easy to train new AI models for every language that is currently known [Jauhiainen et al. 2019b]. Further, some documents are written in more than one language which makes difficult to separate them [Jauhiainen et al. 2019b]. It is necessary to develop methods capable of considering as many languages as possible regardless of the amount of data available. Moreover, it is important to evaluate how well the classifiers perform in terms of the number of languages they are supposed to classify [Jauhiainen et al. 2019b]. The study of a language under a computational perspective usually starts with the application of machine learning models [Linares and Oncevay-Marcos 2017]. Therefore, the investigation and application of

ML models to identify Brazilian indigenous languages is a crucial step to the development of natural language processing tools to such idiomatic groups.

There are two main approaches when dealing with LC [Jauhiainen et al. 2019b]. The first one is supervised language classification, in which, the main objective is to train a model with a labelled corpus, in order to assign a correct label to a newly given text [Jauhiainen et al. 2019b]. One of the most popular methods is textcat [Cavnar et al. 1994], and more recently, deep learning models [Jauhiainen et al. 2019b]. The second approach consists on the application of unsupervised learning, in which the main purpose is to group similar (unlabelled) texts in the same clusters [Jauhiainen et al. 2019b]. In this setting, the well-known K-means algorithm was already used to group text of similar languages [Jauhiainen et al. 2019b].

Therefore, this paper proposes the classification of 26 Brazilian indigenous languages in a supervised and unsupervised manner. To the best of our knowledge, this is the first work to address unsupervised and supervised categorisation of Brazilian indigenous languages. We hope this work to open the field to new research in the field of low resource, indigenous natural language processing, machine translation and language identification. This rest of this paper is structured as follow: Brazilian Indigenous Languages (section 2) , Literature Review (section 3), describes the Materials and Methods (section 4), Experiments (section 5) and Conclusion 6 [1].

2. Brazilian Indigenous Languages

The interest on the proper understanding of Brazilian indigenous languages began with the arrival of the first catholic missionaries in the sixteenth and eighteenth centuries [Severo and Makoni 2014]. The main focus of the Jesuits work was on the grammatisation and the creation of dictionaries and other linguistic materials for teaching purposes [Severo and Makoni 2014]. As quoted by Manoel da Nobrega, it was fundamental to the missionaries to be able to understand and to communicate in the local language [Severo and Makoni 2014]. The documentation process was also influenced by the work of Protestant missionaries in the nineteenth century, specially under the govern of King Pedro II [Severo and Makoni 2014]. The Jesuits produced a large amount of valuable information about Brazilian indigenous languages which might be useful in their study even today. Besides, understanding the beginning of the indigenous languages documentation can bring useful insights about currently available documentation about indigenous languages.

For instance, it is already known that, as a consequence of the documentation and influence of the catholic mission, the indigenous languages suffered the influence of the colonial culture [Severo and Makoni 2014, Fleming 2009]. The indigenous languages received new words and new kinds of linguistic discourses that did not exist in their original culture [Severo and Makoni 2014, Fleming 2009], like confessions, plays, and lyrics. Therefore, despite being a non-European language, the indigenous language might carry much of its characteristics and discourses [Severo and Makoni 2014, Brüzzi 1967]. Examples of created words and new meaning attribution are the words *tupãoka* (*Tupã*+*Oka*), which means house of God (church), and the word *Anhangá rata*, that means devil [Severo and Makoni 2014]. Although, it promotes the enrichment of

[1]Our code and dataset is available at: https://github.com/Tiagoblima/brazilian-languages-eniac2021.git

Linguistic Family	Linguistic unit or dialect	Matthew Chapter 1, Verse 1
Karíb	Macushi	”Pena Abraão wanîpî. Mîîkîrî wanîpî Isaque yu...”
Arawá	Paumarí	”Isaque kaabi’i ada Abraão kohana. Jacó kaabi’i...”
Aruák (Maipure)	Apurinã (Ipurinã)	”Ininiã ia atoko itxa: • Kitxakapirîka Apraão, ...”
	Paresí (Aritî, Haliti)	”Abraão atyo Isaque kaisani, Isaque atyo Jacó k...”
	Teréna (Tereno)	”Eneponeko Âbraum, há’ane neko Izáki. Kene Izák...”
Guaikurú	Kadiwéu	”Abraão jiijaa eliodi Isaque, Isaque jiijaa...”
Jê	Apinayé	”• N Apraãw kra na pre kêp Ijak. • N Ijak kra...”
	Kaingáng	”Abraão v Isaque han. K Isaque v Jacó han. K...”
	Kayapó	”Ingêt Abraão ne Idjak dji. Idjak dji nhym arm...”
	Xavánte	”Abra’ã hã Izatihi mama. Izati hã Zacoho mama. ...”
Karajá	Karajá	”Ibutumy ilabiebohonimy arelyykre. Juhuu tybyni...”
Karíb	Apalaí	”Aparão mokyro Izake zumy. Izake mokyro Jako zu...”
	Bakairí	”Saguhoem kuru ise Jesus idamudo kâengatuly, Ab...
Makú	Nadëb (Guariba, Xiruai)	”Abaraãm taah, Isak. Isak taah, Jakóh. Jakóh ...”
Mawé	Mawé (Sateré-Mawé)	”PIAT EWY Pyno atiatusetpehik teran mesuwe A...”
Maxakalí	Maxakalí	”’Ãmanãm te ’Iyak mũg tak, ha ’Iyak te Yako mũg...”
Mundurukú	Mundurukú	”Isaque ebay Abraão osunuy. Jacó ebay Isaque o’...”
Nambikwára	Nambikwára	”Nxa²ha¹te¹ A³bra³ãu²ah³lai²na² sa²kx...”
Rikbaktsá	Rikbaktsá	”Tapara Abarão niy. Iwaze Isake ta Abarão tse. ...”
Tukano	Tukano	”Abraão Isaque pak niîk niîwĩ. Isaque k’ra J...”
Tupí-Guaraní	Kaapór (Urubu-Kaapór)	”Abrahampa chulinmi Isaac. • Isaacpa chulinñata...”
	Guajajára (dialect Tenetehára)	”Heta Àmàrààw tayr Izak her mae izupe ae. Iz...”
	Guaraní	”Abraão ray ma Isaque, Isaque ray ma Jacó, Ja...”
	Kagwahiva-Tenharím (dialect Kawahíb)	”Ymyahũ Abraãova’ea Isaqueva’ea po’ria hako. Is...”
	Kaiwá (dialect Guaraní)	”Yma ete vaekwe oiko vaekwe Abraão amyrĩ. Ta...”
	Kayabí	”Abraão ga Isaki ga ruwa. Isaki ga Jako ga ...”

Table 1. The table shows the linguistic family of the languages used in the experiments of this paper. It was based on the information available at [Moore et al. 2008]

.

the indigenous languages with more words and meanings, which may not be well suited under its original cultural context [Severo and Makoni 2014]. On the other hand, indigenous languages, such as Tupi, also influenced the use of Portuguese in Brazil by “lending” words and adapting them from the original Tupi [Bisol and Brescancini 2021]. Therefore, throughout the history, Brazilian indigenous languages absorbed different aspects of the European Portuguese and Brazilian Portuguese in it discourse and vocabulary.

Brazilian indigenous languages in general presents wide linguistics variation even inside the same language. On the other hand, some other languages are, in fact, different dialects, given its linguistics proximity as well as of the people that speak them [Moore et al. 2008]. One example is the language spoken by the people of Gavião de Rondônia and their neighbours Zoró [Moore et al. 2008]. Both of them are classified as distinct languages, however they can also be seen as different dialects [Moore et al. 2008]. One reason for this is that there is no systematic documentation of the number of languages [Moore et al. 2008]. It is difficult to affirm how much of them still need to be documented. Also, because of the high similarity of some languages, despite many say that there are 180 languages, in fact, the number might be less than 150 [Moore et al. 2008]. Hence, it is necessary to investigate ways to better analyse data from the indigenous resources at hand, not only to group, but also to verify the linguistic variation inside the same language.

Knowing the exact number of Brazilian indigenous languages is not the only prob-

lem faced by the linguists [Moore et al. 2008]. Another relevant aspect of such studies is the language transmission aspect [Moore et al. 2008]. The transmission rates might indicate the capacity of the language be transmitted from the parents to their descendants and use of the language in a daily basis [Bisol and Brescancini 2021]. High transmission rates corresponds to higher levels of vitality which means languages that are passed from generation to generation and spoken by all age groups fluently [Bisol and Brescancini 2021]. On the other hand, lower transmission rates might indicate that the languages is not spoken by the younger members of the community, or that the children only speak in a specific context. The languages are considered endangered if the it is spoken only by the parents and grandparents [Bisol and Brescancini 2021]. In the work [Moore et al. 2008], from the 150 documented languages, 21% suffer the risk to extinct in a short term due to transmission issues.

Therefore, alongside the history the indigenous languages, its study and documentation have contributed to current knowledge about the languages and indigenous culture. However, there is still a limited amount of research on automatic identification as well as to study the relationship between different indigenous languages and determine its dialects and variants. Hence, throughout this paper, it will be discussed the use of well-known methods and techniques to classify 26 Brazilian indigenous languages and provide insights about the relationship between them.

3. Literature Review

Language identification has been explored throughout many years. The models used for this task vary from simple character n-grams frequencies count, distance measuring algorithms such as textcat to machine learning and deep learning models [Cavnar et al. 1994, Jauhiainen et al. 2019a].

The first step in the use of machine learning methods to language identification tasks is to extract relevant features from texts. Term Frequency (TF) and Inverse Term Frequency (IDF) are well-known methods to such tasks [Dadgar et al. 2016, Kadhim 2019, Li and Shen 2017]. Term Frequency calculates how many times a specific term appears in a document [Yamamoto and Church 2001]. IDF calculates the inverse term frequency of each term considering all the documents [Yamamoto and Church 2001]. Thus, it tells us how informative is each term in the set of documents [Yamamoto and Church 2001].

$$IDF = -log_2 \frac{df(t)}{D} \tag{1}$$

Here, D is the number of documents and df calculates the frequency of the term t in the document d. Therefore, TF-IDF is a simple and effective method to extract valuable information from the text corpus.

The method proposed by [Gebre et al. 2013] found that TF-IDF features of unigrams and bi-grams provided the best results when used as input for the algorithms SVM and Logistic Regression. The SVM had a slightly better performance than Logistic Regression, with an overall 84.55% accuracy against 84.45% of the latter.

The Multinomial Naive Bayes (MNB) algorithm has also been successfully used in many different language classification applications [Jauhiainen et al. 2019b, Bhattu and Ravi 2015, Tan et al. 2014]. In the MNB, each feature is used to calculate

the probability of a text to be written in one of languages to be predicted. In the work by Tan et al. 2014, the algorithm achieved overall 99% accuracy with 5-grams characters, in the task of discriminating between six groups of languages [Tan et al. 2014].

Linear Regression (LR) algorithms usually presents similar results in language classification [Çöltekin and Rama 2016]. It had a slightly worse performance against a Linear model in the works [Çöltekin and Rama 2016, Gebre et al. 2013]. Therefore, it is also computationally efficient to use LR on the language classification tasks, since it can be faster than SVMs.

Unsupervised classification is mostly used when there is no labelled data available for training, and is useful to separate data in groups that have similar features. The K-means algorithm is one the most popular methods in text clustering analysis [Xiong et al. 2016]. In each iteration, k elements (i.e., cluster centres) are selected and the distance between the each element and the other elements is calculated. After that, the cluster centres are updated, and the whole process starts over again [Xiong et al. 2016]. In [Xiong et al. 2016], the K-means algorithm was used to separate text in different topics, such as Economy, arts and sports. The k-means algorithm was used also in the development of methods to language classification, both considering character level and word level n-grams [Amine et al. 2010, Wan 2016]. It is also the scenario considered in this paper what makes the method worth of being investigated in this research.

Handling resources constrains still a significant issue that has been explored in different techniques nowadays. Morphological segmentation can increase the corpus of a set of rare languages. Nevertheless, it needed prior knowledge of the language from a specialist [Kann et al. 2018]. Furthermore, language differs in syntax and morphology. Consequently, it is necessary to develop distinct rules to extract valuable information. As proof, in polysynthetic languages, words are the concatenation of two or more other words [Kann et al. 2018]. Therefore, it is still difficult to increase the corpus of rare language using analyses techniques [Jauhiainen et al. 2019b, Selamat and Akosu 2016, Malmasi et al. 2015]. The language identification of a specific idiomatic group was part of research of the work [Linares and Oncevay-Marcos 2017]. In the work, they used supervised machine learning models to classify 16 Peruvian languages. The technique used to overcome the low textual resource available for the languages, was character's bigrams and trigrams, using TF-IDF as feature extraction method [Linares and Oncevay-Marcos 2017]. Therefore, the exploration of cases where there are low textual resource available is still a significant issue specially for specific idiomatic linguistic groups such as the Peruvian, Brazilian indigenous languages and others.

In summary, in this work, we will explore the use of supervised and unsupervised algorithms to classify 26 Brazilian indigenous languages. Those languages do not have much textual resource available online and therefore, we propose the development of automatic systems that might help linguists on the study and documentation of Brazilian indigenous languages.

4. Materials and Methods

This paper explores the use of supervised and unsuppervised approaches to study

Brazillian indigenous languages. In the supervised setting, we consider the **SVC (Support Vector Classifier)**, **Multinomial Naïve Bayes** and **Logistic Regression** algorithms, in a indigenous language classification task.

Furthermore, this work will use the parallel corpus of the New Testament of the Bible of 26 different indigenous language and the Portuguese. To the best of our knowledge, this is the first study to apply machine learning methods to this corpus.

4.1. Multi-Language Text Corpus

Evaluating techniques using parallel corpora are meaningful but it is difficult to find significant parallel corpus between languages [Jauhiainen et al. 2019b]. It would be helpful to fairly evaluate techniques and decide which one is the best. To guarantee a precise evaluation, it is interesting to use the same text corpora [Jauhiainen et al. 2019b]. The experiments in this work consider a corpora composed of verses from the New Testament translation of 26 Brazilian indigenous languages [Angelo 2016]. Standard pre-processing steps were applied, such as the removal of HTML tags, punctuation symbols and cross references. A total of 100 verses from each language were chosen and split into train and test documents. Table 2 presents a sample of the same verse in all 26 languages considered in this study. Therefore, this paper will use the corpus from the New Testament of the Bible to supervised and unsupervised language identification of the Brazilian indigenous languages.

4.2. Evaluation Metrics

We chose two metrics for the supervised classification and two metrics for unsupervised classification. For supervised classification the metric, the first metric considered was accuracy that gives an overview of the performance of each classifier [Pedregosa et al. 2011]. Despite its simplicity, the accuracy has been used in other works of language classification to measure the performance of different classifiers [Zampieri et al. 2015]. Secondly, the F1 score was used to have a deep view of the results provided by each classifiers [Pedregosa et al. 2011]. Different from accuracy, the f1-score provides not only how many of the labels were predicted correctly by the classifier but also balances the counts of true positive and true negative generated by the classifier output [Pedregosa et al. 2011]. Besides, to evaluate the quality of the clusters formed by K-means algorithm, we chose the Davies-Bouldin score and silhouette coefficient score metrics. The first metric, Davies-Bouldin score, measures the similarity between two clusters considering the distance within each clusters and between the clusters [Pedregosa et al. 2011] and the lower values indicate better results. On the other hand, silhouette coefficient considers the distance between the examples within the clusters and between the nearest cluster [Pedregosa et al. 2011], the best value is 1 and the worse is -1.

5. Experiments

Firstly, we use TF-IDF transformation with words bigrams and trigrams to extract features from the original 2600 verses (100 from each language). The used algorithms are available at the scikit-learn python library [Pedregosa et al. 2011, Buitinck et al. 2013]. The proposal of using bigrams and trigrams was also explored in the work [Linares and Oncevay-Marcos 2017]. A visualisation of a 2-D t-SNE [LJPvd and Hinton 2008, Krijthe and Van der Maaten 2015, Van Der Maaten 2014]

projection of the extracted TF-IDF from 100 samples (per language) can be seen in figure 1. Each point corresponds to a verse (i.e., a document) from the dataset. One can note that the TF-IDF extracted features can provide a highly separable representation of the languages. It is clear to see that there are a close relationship between each language when word n-grams is used to TF-IDF feature extraction. Besides, when using character analyzer, despite a considerable number of languages, all of them are well separate with almost no overlaps what turns the classification problem much easier.

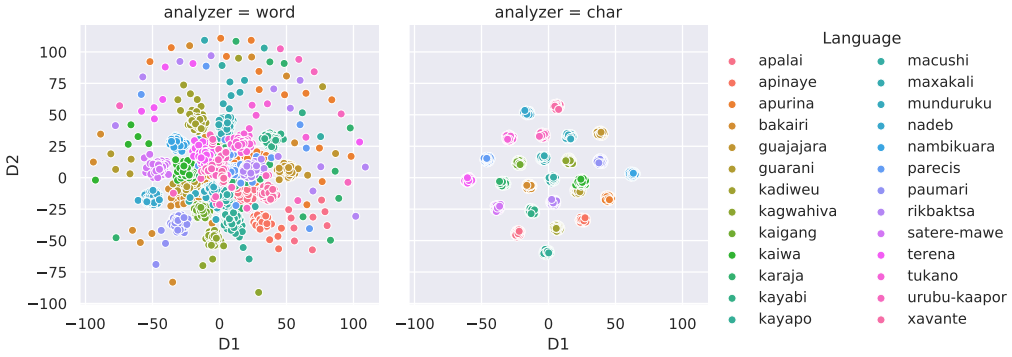

Figure 1. The TSNE projection of the the languages.

5.1. Supervised Classification

In the first experiment, using the supervised classifiers exposed at table 3, to verify the performance of different classifiers considering only word bigrams and trigrams. It was used 10 times StratifiedKFold to mantain the percentagem of labels as it is in the original dataset [Pedregosa et al. 2011]. The LinearSVC classifier shows a superior performance than the other classifiers.

When character n-grams are considered, the results is improved for all the algorithms what confirms the initial hypothesis that character n-grams provided a much better separation of the languages than word n-grams as it appears at the figure 1. Further, the same experiment was made considering different sizes of dataset from 20% to 100% of the samples dataset of 26000 examples. The figure 2 shows the consistence of the TF-IDF with character bigrams and trigrams method on classifying indigenous languages.

5.2. Unsupervised Classification

In the unsupervised experiments, it was tested different numbers of clusters considering numbers from 2 to 30. We also extract bigrams and trigrams features now considering the whole dataset of 2600 verses. The two different metrics shows that character n-grams provide much better clusters than word n-grams.

The first metric, Davies-Bouldin score, shows much lower values when considering n-grams character what means much better results. When the initial number of clusters are considered (it is two), the lowest value is close to the original number of languages that is 26. In the second metric, there is a similar outcome, the metrics hits it is best value when it is close to the real number of languages evaluated.

It stands out that combining both metrics might provide a exactly picture of the real number of languages when character n-grams are used.

Linguistic unit or dialect	Top-10 TF-IDF Tokens
Apalaí	xine, toto, poko, ritonõpo, ase, pyra, tykase, jezu, mana, eya
Apinayé	m, n, amnhĩ, pa, kot, nhũm, kãm, ri, hã, na
Apurinã (Ipurinã)	teoso, iua, xesosi, ininiã, ninoa, hĩte, nota, sãkire, maerekati, kotxi
Bakairí	jesus, deus, modo, myani, urâ, warâ, wâgâ, kely, aguely, ise
Guajajára (dialect Tenetehára)	ae, nehe, wà, mae, ihe, rehe, putar, kury, kwaw, pe
Guaraní	hae, vae, kuery, ma, rami, vy, e, rã, aguã, rupi
Kadiwéu	ni, me, ane, odaa, ica, oji, aneotedo, odi, ijo, jesus
Kagwahiva-Tenharím (dialect Kawahíb)	ga, pe, ei, ti, ea, ji, tupana, po, jesus, neh
Kaigáng	tóg, ag, ti, t, mũ, k, m, nĩ, ãjag, g
Kaiwá (dialect Guaraní)	vae, xe, hei, vy, pe, kwéry, íxupe, hesu, nhandejáry, ramo
Karajá	tahe, heka, kia, deuxu, jesuisi, tii, iny, mahãdu, rare, bede
Kayabí	ga, gã, je, jau, nupe, futat, upe, mamae, aeramũ, ã
Kayapó	ne, me, ar, nhym, arm, kam, kum, ba, kute, dja
Macushi	pe, mîrîrî, pî, pra, to, moropai, paapa, tapî, tîîse, morî
Maxakalí	tu, te, ha, ãktux, topa, ax, yõg, hãm, ũn, tikmũ
Mundurukú	ip, io, deus, jesus, ma, kay, be, õn, pe, soat
Nadëb (Guariba, Xiruai)	do, naa, bë, ti, h, ta, bä, m, hã, doo
Nambikwára	la^{2}, a^{2}, na^{2}, nxe^{3}, txa^{2}, $ĩ^{3}$, h^{1}, ta^{1}, su^{2}, $hxai^{2}$
Paresí (Aritĩ, Halíti)	hoka, atyo, nexa, jesus, kakoa, maisa, enore, hatyo, waiyexe, hiyeta
Paumarí	ida, ra, aha, va, ora, bana, ari, ada, ni, deus
Rikbaktsá	deus, niy, bo, sesus, batu, humo, ty, kytsa, my, zeka
(Mawé) Sateré-Mawé	hap, yt, hawyi, mii, tupana, pe, waku, rayn, kahato, iesui
Teréna (Tereno)	ne, oviti, itukó, jesus, xâne, ákoti, koati, vo, oku, enepone
Tukano	k, y, niîwĩ, niî, h, âk, tohô, bu, re, jesu
(Kaapór) Urubu-Kaapór	mana, cay, nil, ya, sumä, chay, tayta, niptinmi, allin, manam
Xavánte	te, za, wa, re, hã, ra, mono, ma, na, aba

Table 2. The table shows the top 10 tokens for each language according to TF-IDF.

5.3. Discussion

TF-IDF is a powerful method to extract features from textual data and the particular analyzes of word or character much make a huge difference in the final outcome result. First, the figure 1 shows how words features might be close related between different languages when analyzing words. The character analyze, however, provides a much better separation of the languages. As a result, in our experiments, the simple change of the feature analyzed could improve the result of all the classifiers considered in the supervised approach. It might help the classification of the language that are already known such as the ones that we considered in this paper.

Secondly, the unsupervised experiments showed that character n-grams were ideal to take the real number or close to the real number of languages in the dataset when the correct metrics are used and well combined. It happens because the metrics achieves their optimal result when the number of languages is exactly or close to the real number of languages. It might mean that clusters are well separated from each other and with high level of homogeneity. In contrast, when the word analyzes is considered, the k-means performance is unstable and unpredictable. It makes the detection of the ideal number of clusters (languages) much harder in the case of Davies-Bouldin metrics. When the silhouette coefficient score is considered the model performance shows small changes through the tests what can make the detection of optimal difficult. Since it was considered languages from the same linguistic family, it might also be helpful to separate dialects or language that belong to the same family. Those conclusion might help in the development

	accuracy			f1_macro		
	mean	median	std	mean	median	std
classifier_name						
Char TFIDF+LinearSVC	1.000	1.000	0.000	1.000	1.000	0.000
Char TFIDF+LogisticRegression	1.000	1.000	0.000	1.000	1.000	0.000
Char TFIDF+MultinomialNB	1.000	1.000	0.000	1.000	1.000	0.000
Word TFIDF+LinearSVC	0.972	0.979	0.018	0.975	0.980	0.015
Word TFIDF+LogisticRegression	0.972	0.977	0.017	0.974	0.978	0.015
Word TFIDF+MultinomialNB	0.950	0.954	0.018	0.953	0.956	0.015

Table 3. The table shows the classification performance of Logistic Regression, LinearSVC and MultinomialNB

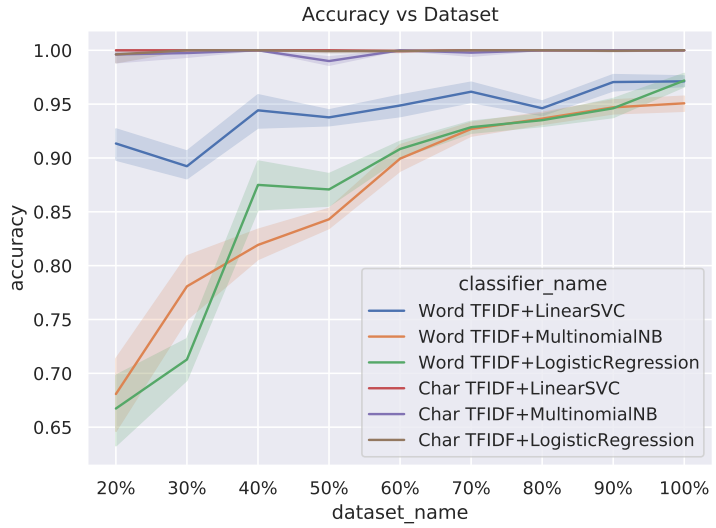

Figure 2. Showing accuracy considering different sizes of dataset. It starts with only 20% of the samples and ends with 100% of the 26000 examples.

of application to detect the right number of Brazilian indigenous languages in a collection of text as long as it is one the current problems when dealing with Brazilian indigenous languages [Moore et al. 2008].

There are also other application of the current finds. One of them is the detection of a unknown or unseen language as proposed by [Jauhiainen et al. 2019b]. There are different attempts to address this problem as exposed by [Jauhiainen et al. 2019b]. An unsupervised approach can be used along with a supervised approach to detected new languages. Finally, in this first experiments using Brazilian indigenous languages, character n-grams can make a huge difference on language classification. More studies much show the application of this method and its efficiency considering other languages and scenarios.

6. Conclusion

This work explored some Brazilian indigenous languages digital resources, and proposed the application of machine learning methods and models to help their analysis and stud-

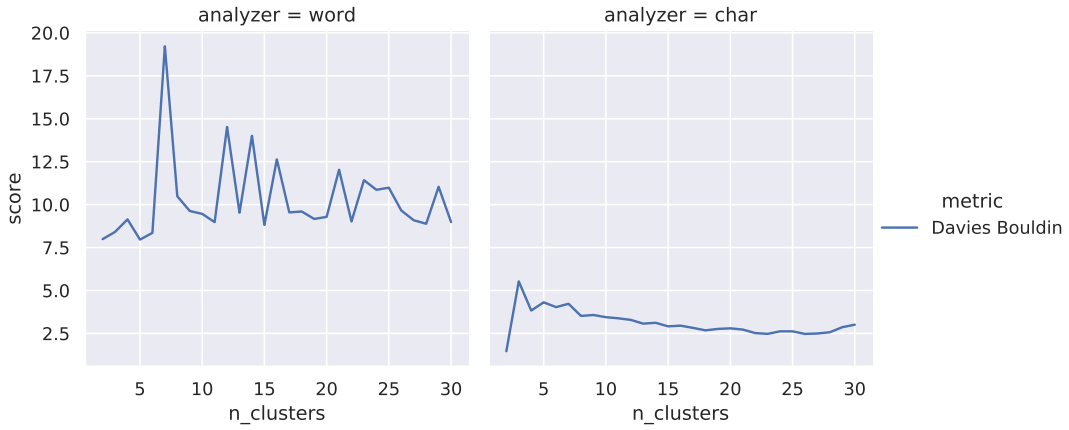

Figure 3. Using Davies Bouldin score to measure the cluster quality.

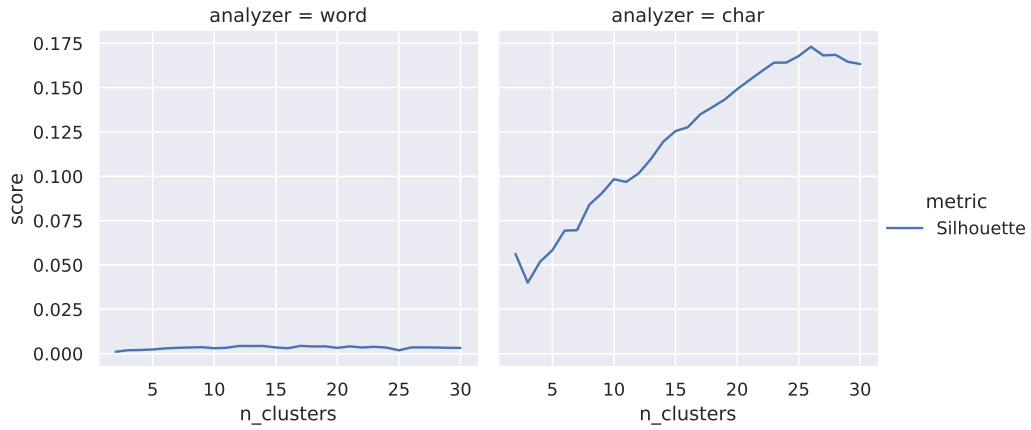

Figure 4. Using Silhouette coefficient score to measure the cluster quality.

ies. The work is motivated by the lack of scalable methods to analyse digital content in such low resource languages, in order to aid the documentation of such endangered languages. In this context, language classification might have a significant role on the documentation of those languages, as well as to study similarities among dialects. Our research pointed that overall, they do not share a large common vocabulary or grammatical structure. Further, we expect that this work may encourage other researchers to use machine learning models to this domain, as well to the development of digital solutions on such low resource languages, such as Automatic Machine Translation.

References

Amine, A., Elberrichi, Z., and Simonet, M. (2010). Automatic language identification: An alternative unsupervised approach using a new hybrid algorithm. *Int. J. Comput. Sci. Appl.*, 7(1):94–107.

Angelo (2016). 26 versões da bíblia em idiomas indígenas para mysword.

Bhattu, S. N. and Ravi, V. (2015). Language identification in mixed script social media text. In *Fire workshops*, pages 37–39.

Bisol, L. and Brescancini, C. R. (2021). *Contemporary phonology in Brazil*. Cambridge Scholars Publishing.

Brüzzi, A. A. d. S. (1967). Observações gramaticais da língua daxseyé ou tucano. *Centro de Pesquisas de Iauaretê*.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer.

Çöltekin, Ç. and Rama, T. (2016). Discriminating similar languages with linear svms and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24.

Dadgar, S. M. H., Araghi, M. S., and Farahani, M. M. (2016). A novel text mining approach based on tf-idf and support vector machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116.

Drude, S., Jr, N. G., and Galucio, A. V. (2007). Avanços da documentação sobre línguas indígenas no Brasil. page 4.

Fleming, L. (2009). Indigenous language literacies of the northwest amazon. *Working Papers in Educational Linguistics (WPEL)*, 24(1):3.

Gebre, B. G., Zampieri, M., Wittenburg, P., and Heskes, T. (2013). Improving native language identification with tf-idf weighting. In *the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, pages 216–223.

Jauhiainen, T., Lindén, K., and Jauhiainen, H. (2019a). Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Jauhiainen, T. S., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019b). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Kadhim, A. I. (2019). Term weighting for feature extraction on twitter: A comparison between bm25 and tf-idf. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pages 124–128.

Kann, K., Mager, M., Meza-Ruiz, I., and Schütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

Krijthe, J. H. and Van der Maaten, L. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation. *R package version 0.13, URL https://github.com/jkrijthe/Rtsne*.

Li, Y. and Shen, B. (2017). Research on sentiment analysis of microblogging based on lsa and tf-idf. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pages 2584–2588.

Linares, A. E. and Oncevay-Marcos, A. (2017). A low-resourced peruvian language identification model. In *CEUR Workshop Proceedings*. CEUR-WS.

LJPvd, M. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *J Mach Learn Res*, 9:2579–2605.

Malmasi, S., Dras, M., et al. (2015). Automatic language identification for persian and dari texts. In *Proceedings of PACLING*, pages 59–64.

Moore, D. and Galucio, A. V. (2016). 2. perspectives for the documentation of indigenous languages in brazil. In *Language documentation and revitalization in Latin American contexts*, pages 29–58. De Gruyter Mouton.

Moore, D., Galucio, A. V., and Gabas Jr, N. (2008). O desafio de documentar e preservar as línguas amazônicas. *Scientific American Brasil*, 3:36–43.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Selamat, A. and Akosu, N. (2016). Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University-Computer and Information Sciences*, 28(4):457–469.

Severo, C. G. and Makoni, S. B. (2014). Discourses of language in colonial and postcolonial brazil. *Language & Communication*, 34:95–104.

Tan, L., Zampieri, M., Ljubešic, N., and Tiedemann, J. (2014). Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15. Citeseer.

Van Der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245.

Wan, A. (2016). Leveraging data-driven methods in word-level language identification for a multilingual alpine heritage corpus. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 45–54.

Xiong, C., Hua, Z., Lv, K., and Li, X. (2016). An improved k-means text clustering algorithm by optimizing initial cluster centers. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 265–268. IEEE.

Yamamoto, M. and Church, K. W. (2001). Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.

Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Nakov, P. (2015). Overview of the dsl shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9.