Mapping Ancestry through Surnames: Machine Learning Approaches Applied to Brazilian Data

Arthur Lins Wolmer[1], Diego de Freitas Bezerra[1]

CESAR School, Recife, Pernambuco, Brazil
{alw, dfb2}@cesar.school

Abstract. The classification of surname origin as a proxy for ethnic background estimation has long supported sociological, demographic, and genetic studies, particularly in countries with diverse migratory histories. In this article, we introduce a new Brazilian dataset constructed from over one million historical immigration records, propose a pipeline for surname extraction and disambiguation, and evaluate multiple supervised classifiers based on character-level n-grams. In addition to replicating classical models, we implement graph-based methods and an ensemble classifier. Our results confirm the competitiveness of traditional approaches while achieving significant gains with the ensemble model.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Supervised learning by classification**; *Ensemble methods*; *Classification and regression trees*.

Keywords: ancestry inference, name disambiguation, ensemble learning, graph-based models, surname classification

1. INTRODUCTION

Discussions about racial and ethnic inequalities have resurfaced in Brazil, particularly in light of debates over affirmative action policies, systemic racism, and social mobility [Heringer, 2024; Ribeiro and Carvalhaes, 2024]. However, these discussions are frequently limited to the five official ethnic background categories defined by the Brazilian Institute of Geography and Statistics (IBGE): white, mixed/brown, black, East Asian, and Indigenous. Although useful, this classification fails to capture the complex and diverse ethnic ancestry of the Brazilian population, a country shaped by centuries of immigration from Africa, Europe, the Middle East, Asia, and other regions.

To the best of our knowledge, no regular surveys are carried out by IBGE or any other public or private institution to gather data on the ethnic ancestry of the Brazilian population. Schwartzmann [1999] and a study carried out by IBGE in 2008 [IBGE, 2011] presented some insights in this matter, but these were not repeated later nor broadly representative of the entire national population.

This gap persists despite the fact that Brazil was the third country in the Western Hemisphere in number of immigrants received between 1820 and 1960, totaling approximately 5.4 million individuals from various historical and social backgrounds, both among themselves and in relation to the population already living in the country [CPDOC, 2000]. This fact supports the hypothesis that persistent structural inequalities may exist among descendants of these different groups.

In this context, Monasterio [2016] proposed the use of machine learning-based surname classification as a tool to estimate ethnic ancestry in social research. His study introduced a methodology to classify surname-based ancestry in Brazil, leveraging historical surname databases and applying fuzzy string matching and machine learning to a large-scale administrative dataset, the 2013 edition of the Relação

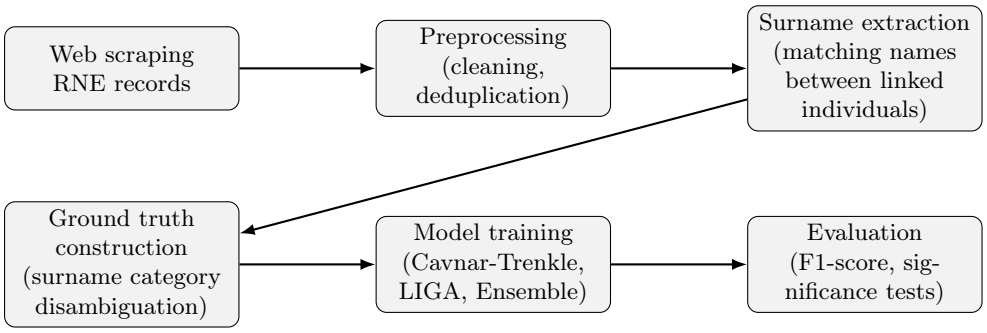

Fig. 1. Pipeline for the assessment of surname classification approaches

Anual de Informações Sociais (RAIS), published by the Brazilian Ministry of Labor and Employment[1].

We propose a surname-based ancestry classification method related to that presented by Monasterio [2016]. While his study focused on five ancestry groups and relied on fuzzy matching combined with simple classifiers, our work presents the following contributions:

—Firstly, we include an additional ancestry category to better reflect Brazil's immigration history;
—Secondly, we construct a new dataset by scraping and processing over 1,000,000 records from the National Registry of Foreigners, improving data specificity and transparency;
—Thirdly, we present a pipeline for data cleaning, surname extraction, and ambiguity resolution;
—Finally, we explore a broader set of machine learning techniques, including graph-based models and an ensemble classifier, and evaluate them using statistical tests and macro F1-scores to assess performance across imbalanced classes.

This article is organized as follows: Section 2 describes the data and the process employed into our evaluation; Section 3 presents the obtained results; At last, Section 4 concludes our work and delineates some future contributions.

2. DATA AND METHOD

This section describes the dataset construction process and the machine learning methods employed in the study. The ancestry classification pipeline consists of several steps, beginning with the extraction and cleaning of historical records, then building a reliable surname-to-ancestry mapping, and finally training and evaluating predictive models. All scripts, models, and processed data used in this pipeline are publicly available in our GitHub repository.[2]. Figure 1 provides an overview of the entire process and serves as a roadmap for the components discussed in the subsequent subsections.

2.1 Data

The dataset used to map surnames to ethnic ancestries was built from entries in the National Registry of Foreigners (Registro Nacional de Estrangeiros – RNE), preserved at the Public Archive of the State of São Paulo within the collection titled Delegacia de Estrangeiros. Established in 1938 during Brazil's Estado Novo regime, the RNE was mandatory for all foreign residents in the country under the age of 60[3]. Although it is limited to São Paulo state, there is, to the best of our knowledge, no larger or more diverse (in terms of nationalities represented) dataset on foreign individuals in Brazil.

[1]Available at: `http://www.rais.gov.br/sitio/index.jsf`. Accessed on: June 13, 2025.
[2]`https://github.com/alwolmer/ngram_surname_classifiers/`
[3]Available at: `https://www.arquivoestado.sp.gov.br/web/acervo/solicitacao_certidoes/delegacia`. Accessed on June 4, 2025.

Table I. Ancestry categories in the scraped dataset

Category	Nationalities	Number of records
Iberian	Portuguese, Spanish	466,535
Italian	Italian	233,830
Japanese	Japanese	206,411
Germanic	German	71,570
Arabic	Syrian, Lebanese	33,309
Eastern European	Polish, Russian, Ukrainian	19,487

The records available online correspond to indexes of archived administrative cases and contain the following fields: full name of the foreigner, marital status, registry or certificate number, nationality, father's name, mother's name, and year of arrival in Brazil.

We scraped the records using a fail-safe Python script on a virtual machine to extract all entries pertaining to the nationalities that represent the ethno-linguistic groups most prevalent among 19th and 20th century immigrants to Brazil, according to official statistics tabulated in IBGE [2007] (p. 226). The scraping process generated a total of 1,058,880 raw records. Table I presents the ancestry classes, corresponding nationalities, and record counts.

2.1.1 *Preprocessing.* The collected data was preprocessed in order to extract a set of highly reliable surname-ancestry pairs. Many records contain missing values, particularly in the registry/certificate number and parents' names fields. Manual inspection revealed duplicate individuals (i.e., multiple entries for the same person), frequently with discrepancies in the recorded year of arrival.

To address these issues, the dataset was cleaned through three steps: (i) normalizing the name fields (for the immigrant and both parents) by converting all text to lowercase and all characters to their ASCII counterparts; (ii) removing entries with numbers, punctuation, or placeholder expressions such as "desconhecido(a)" (unknown) or "sem informação" (no information); and (iii) eliminating records with incomplete or single-word names. After cleaning, 869,304 records were kept.

Deduplication was then performed by identifying (i) records sharing the same registry number and full name or (ii) those with identical combinations of full name, father's name, and mother's name.

When removing duplicates, we prioritized entries with more complete parental names (sometimes abbreviated) and older arrival dates. The resulting dataset had 726,124 deduplicated records, each with at least two valid names across the three primary fields (full name, father, and mother).

2.1.2 *Surname extraction.* Using all three name fields, we used a heuristic to extract hereditary surnames. Assuming that shared surnames between related individuals (either by descent or marriage) are more likely to appear at the end of the full name, we used a metric which we dubbed *endness* — the sum of the position indices for a shared word between two names — to obtain the last shared common surnames between the three possible pairs for each data entry:

—**Parents' surname** (`parent_surname`): last common word between father and mother;
—**Father's surname** (`father_surname`): last common word between the immigrant and the father;
—**Mother's surname** (`mother_surname`): last common word between the immigrant and the mother.

Extraction was not always possible, as it depended on name overlap in the pairwse comparisons. Nevertheless, this approach effectively filtered out many noisy or misleading cases, such as:

—Names changed due to marriage (foreign women married to people of other origins);
—Reversed name order (as occasionally occurs with Japanese, and to a lesser extent, Italian names);
—Non-hereditary elements that follow given names (e.g., "filho", "júnior", "sobrinho").

Table II. Unique surnames with final category assignment

Category	Entries	Unique surnames
Italian	146,527	33,426
Iberian	323,747	18,987
Germanic	40,641	15,638
Japanese	138,984	13,172
Eastern European	12,222	7,354
Arabic	23,241	5,873

At the end of the process, 94.39% of the records retained at least one of the three extracted surnames (covering up to 98% of Japanese individuals, at best, and 72% of Ukrainian nationals at worst). We discarded records for which no surname could be extracted, keeping a total of 685,362 entries. All unique values between parent, father, and mother surnames from each of the dataset's retained entries were tabulated, resulting in 95,373 unique surnames, mapped to a category and with a count attribute relative to the amount of tuples where it appeared. In our pipeline, it was assumed that a surname that appeared in only one category belonged to it. However, we did not remove ambiguous surnames i.e. those that appeared in more than one ancestry category.

For each ambiguous surname, we used a heuristic consisting of three steps: first, we calculated the relative frequency of each surname per category to normalize across groups with different sample sizes. Second, we computed the frequency ratio of the most frequent category to each of the others. Third, we calculated ratios between all of the frequency values; if the ratio of a given category's frequency to all others (considered pairwise) equaled or exceeded a decision threshold, the surname was considered predominant. In this step, we arbitrarily set the decision threshold at 2, meaning if a name was at least twice as common in one category compared individually to all others, it would be assigned to it. Otherwise, the surname was excluded from the ground truth.

Using this approach, 3,109 out of 4,032 ambiguous surnames were confidently classified (77%). The final result was a set of 94,450 surname–origin pairs, with no overlaps, distributed as in Table II.

Even when extraction rates are comparable, the number of entries in each ancestry category is not directly proportional to the number of unique surnames obtained. This disparity is due in part to inherent anthroponymic diversity—some ethnic or linguistic groups will simply have more distinct surnames than others in an equal-sized random sample of personal names. It is also influenced by multilingualism within national origins (only Portugal and Japan of the countries under consideration were primarily monolingual during the 19th and 20th centuries), as well as greater spelling variation among languages that differ more from Portuguese or allow for multiple transliteration schemes into Latin. In such cases, a single original surname may appear in several variations, artificially increasing the number of distinct names.

2.2 Method

Machine learning models for surname classification require a *ground truth* dataset labeled by ancestry. As a result, these models are intended to categorize names that have not been previously labeled or observed. In the approach proposed by Monasterio [2016], the classification process followed three sequential stages: first, exact matching with the ground truth was used; second, the remaining names were classified using fuzzy matching with edit distance OSA[4]=1; and third, any residual names were processed using a supervised model trained on the ground truth.

In this article, we focus exclusively on the training and evaluation of supervised models. All models employed are based on the n-gram approach, which relies on decomposing text into sequences of `n`

[4]Optimal String Alignment, which measures the minimum number of single-character operations—insertions, deletions, substitutions, or adjacent transpositions—required to transform one string into another.

Parameter	Description
n	Size of the n-grams.
top_k	Number of most frequent n-grams to include in the profile.
distance	Distance metric used for inference. The following options were implemented:
`rank`	Out-of-place ranking distance, as proposed in Cavnar and Trenkle [1994].
`cosine_tfidf`	Cosine distance computed over TF-IDF vectors.
`match_rate`	Proportion of shared n-grams.

Table III. Parameters and their descriptions for the n-gram profile model.

consecutive characters. For example, the word *machine* can be decomposed into 4-grams such as *mach*, *achi*, *chin*, and *hine*. Our choice of the ngram approach is based on its use in previous Brazilian surname classification studies, allowing us to provide clear and directly comparable performance evaluations of these traditional methods [Monasterio, 2016].

For each approach, we implement a classifier class compatible with the Python *scikit-learn* library API. Given the inherently imbalanced class distribution mentioned in the previous subsection, we considered accuracy to be an insufficient evaluation metric for this task. We used the *F1-score* as the primary performance metric because it gives equal weight to all classes and promotes a balance between precision and recall, better reflecting the model's ability to distinguish between categories.

2.3 N-gram classifiers based on Cavnar-Trenkle

Cavnar and Trenkle [1994] formalized the use of n-gram profiles for text classification. Each class is represented by the top-k most frequent n-grams in the training data. To classify a new entry, its profile is generated and compared to each class profile using a distance metric; the entry is assigned to the closest match [Nelson and Shekaramiz, 2022; Jauhiainen et al., 2019]. The `NGramClassifier` class implements this approach with the configurable hyperparameters listed in Table III.

As a baseline, we replicated the model presented by Monasterio [2016] with `n=3`, `top_k=1250`, and the `rank` distance metric. We also evaluated variants with the same `n` and `k` values but using `cosine_tfidf` and `match_rate` as distance metrics.

2.4 Graph-based classifiers: LIGA and its optimizations

The LIGA (Language Identification, Graph-based Approach) algorithm [Tromp and Pechenizkiy, 2011; Jauhiainen et al., 2019] differs from traditional n-gram profiling methods by modeling not only the frequency of n-grams, but also the ordering in which they appear, capturing sequential structure in a labeled graph representation. It was developed with the express goal of classifying *short and noisy texts*, as analyzing only n-gram frequencies had shown to perform poorly for some such applications.

LIGA models are trained by constructing a weighted, directed multigraph with vertices representing character n-grams and edges capturing their transition patterns in labeled data. During inference, the input string is decomposed into n-gram and traversed the graph to calculate per-class scores based on observed nodes and transitions. These scores are then normalized to account for variation in class size, and the input is assigned to the class with the highest overall score. Our implementation, the `LIGAClassifer` class, includes the baseline algorithm, a configurable `n` parameter, and two optional enhancements from Vogel and Tresner-Kirsch [2012]: Vogel and Tresner-Kirsch [2012]:

—**Log-normalized weights** (`use_log=True`): node and edge frequencies are transformed via $\log(1+\text{count})$ prior to normalization, mitigating the effect of highly frequent but potentially uninformative patterns, consistent with Zipf's law;

—**Median-based scoring** (`use_median=True`): uses the median of node and edge contributions instead of their sum, improving robustness to outliers and dominant n-grams.

2.5 Ensemble Classifier: Combining N-Gram and Graph-Based Features

We implemented the `NgramEnsembleClassifier` class, a composite model that aggregates outputs from both n-gram profile-based and graph-based classifiers derived from the LIGA algorithm, to investigate whether model ensembling could improve robustness and generalization. The goal was to use the complementary strengths of various n-gram distances, lengths, and scoring mechanisms in a single predictive framework. The ensemble is organized in two stages. In the first stage, individual classifiers are trained on the input data. This includes:

—**NGram classifiers**: for each distance metric (`rank`, `cosine_tfidf`, and `match_rate`), we train multiple `NGramClassifier` instances with varying values of `n` and corresponding values of `top_k`, resulting in 15 base models (`n` = 1 through `n` = 5).

—**LIGA classifiers**: we additionally train LIGA models with $n = 3$, using the four different configurations of the two available optimizations: log-normalized weighting and median-based scoring.

Each base classifier produces a per class probability distribution for a given input. These outputs are concatenated to form the feature vector passed to the final-stage classifier. We also include the character length of the input name as a scalar feature. In total, the ensemble model processes 121 features (19 models $\times$ 6 classes + 1 length feature).

In the second stage, a gradient boosting model (`LGBMClassifier` instance) is trained on the extracted features to learn optimal combinations of the base models' outputs. To mitigate the effects of class imbalance we trained the classifier with class-balanced sample weights rather than attempting to generate synthetic surnames (which we judged infeasible for capturing genuine ethno-linguistic variation). At test time, however, we preserve the original proportional distribution of surnames—reflecting real-world prevalence—so that performance metrics remain meaningful.

Because it relies on multiple base classifiers and a second-stage learner, the ensemble approach is significantly more computationally expensive than simpler models. In our experimentation, it took around two orders of magnitude more time per training + inference iteration than the individual `NGramClassifier` instances tested, with the same dataset size and hardware resources.

3. RESULTS

To rigorously assess model performance and determine the statistical significance of differences among classifiers, we conducted 30 repetitions of training and evaluation using stratified splits. In each iteration, we performed 5-fold cross-validation over the training set and also evaluated the results from each model on a held-out 20% test set. The metrics reported here refer to the mean F1-scores across folds (for training performance) and the holdout scores (for direct comparison across models, as the holdout was the same for all models within the same iteration).

We summarize the average macro F1-score and per-class F1-scores for each model in Table IV. Confidence intervals at the 95% level were computed using the standard error of the mean over the 30 repetitions. The ensemble classifier achieved the highest macro F1-score, as well as the best per-class performance in every instance. We note that each class seemed to have some level of intrinsic distinctiveness, and the ranking of performance tended to be, from most easily identifiable to least: Japanese, Italian, Germanic, Iberian, Eastern European, and Arabic.

Figure 2 summarizes the comparative performance of all models by displaying the mean macro F1-scores along with their 95% confidence intervals across 30 cross-validation repetitions. The Ensemble classifier has the highest mean score (0.7708), significantly outperforming the others. Next, the traditional NGram classifier with rank distance performed best (0.7207), followed by the median-enhanced LIGA variant (0.7009). Models that combined log-normalized weights and match-based distances produced lower results, and the unoptimized LIGA baseline had the lowest F1-score (0.5683).

Table IV. Mean F1-scores across 30 repetitions (cross-validation folds); 95% confidence intervals in parentheses

Model	Macro F1	Iberian	Italian	Japanese	Germanic	Arabic	Eastern
Ensemble	**0.7708 (±0.0005)**	**0.7397 (±0.0008)**	**0.8339 (±0.0006)**	**0.9256 (±0.0006)**	**0.7919 (±0.0009)**	**0.6734 (±0.0017)**	**0.6604 (±0.0013)**
NGram (rank)	0.7207 (±0.0006)	0.6774 (±0.0009)	0.8050 (±0.0006)	0.8820 (±0.0007)	0.7597 (±0.0011)	0.5840 (±0.0016)	0.6162 (±0.0017)
LIGA (median)	0.7009 (±0.0006)	0.6561 (±0.0010)	0.7809 (±0.0007)	0.8694 (±0.0007)	0.7475 (±0.0011)	0.5455 (±0.0016)	0.6062 (±0.0018)
NGram (cosine_tfidf)	0.6862 (±0.0007)	0.6583 (±0.0009)	0.7694 (±0.0008)	0.8415 (±0.0009)	0.7402 (±0.0010)	0.5158 (±0.0017)	0.5922 (±0.0017)
LIGA (log+median)	0.6203 (±0.0007)	0.6126 (±0.0010)	0.7019 (±0.0009)	0.7318 (±0.0008)	0.6628 (±0.0015)	0.4221 (±0.0014)	0.5905 (±0.0019)
NGram (match_rate)	0.6049 (±0.0007)	0.5476 (±0.0011)	0.6160 (±0.0012)	0.8167 (±0.0016)	0.6963 (±0.0011)	0.4026 (±0.0012)	0.5498 (±0.0022)
LIGA (log)	0.6026 (±0.0007)	0.6242 (±0.0011)	0.6795 (±0.0011)	0.6876 (±0.0008)	0.6437 (±0.0016)	0.3786 (±0.0012)	0.6017 (±0.0018)
LIGA (plain)	0.5683 (±0.0007)	0.5412 (±0.0012)	0.7157 (±0.0007)	0.6393 (±0.0011)	0.6800 (±0.0012)	0.3683 (±0.0018)	0.4652 (±0.0020)

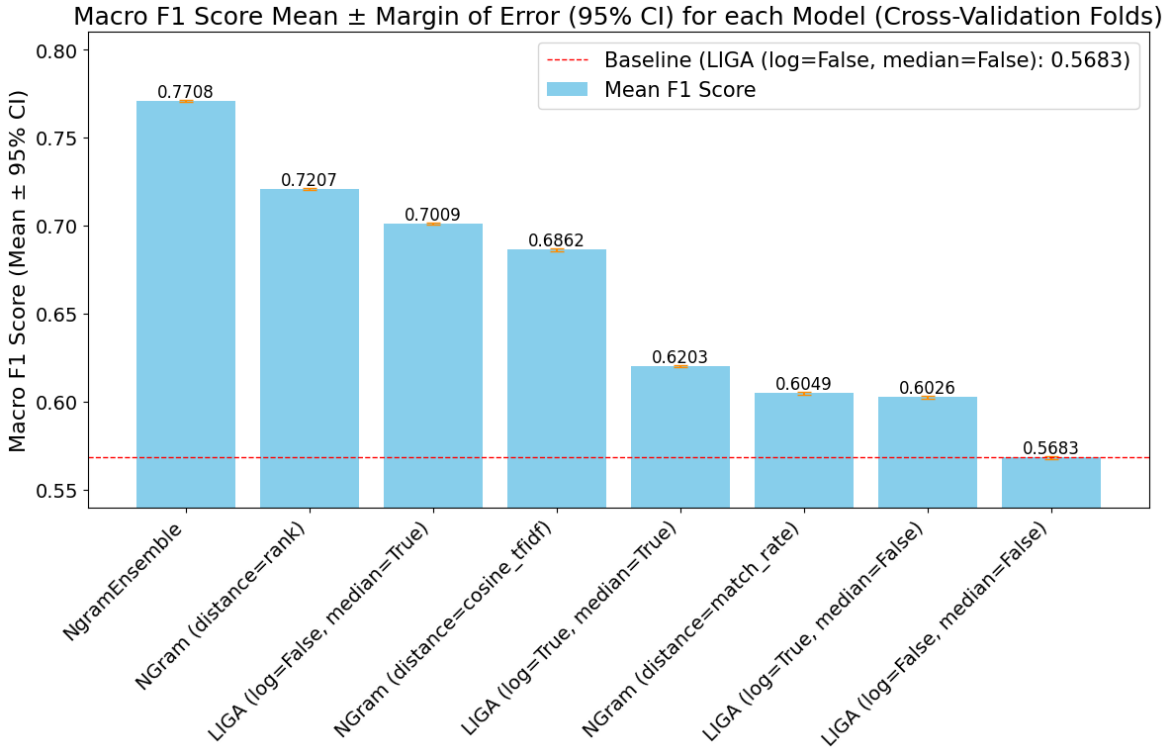

Fig. 2. Mean macro F1-score and 95% confidence intervals across 30 repetitions (cross-validation folds). The red dashed line indicates the baseline performance of the worst model (unoptimized LIGA classifier).

Finally, we used the holdout set scores from each iteration and conducted one-tailed paired t-tests between the Ensemble classifier (the one with the highest mean of F1-Score) and each of the others evaluated. One-tailed tests were chosen because our objective was to confirm that the Ensemble outperforms the base models—rather than merely detect any difference. All tests were conducted with a significance level of $\alpha = 0.05$. The t-statistics and corresponding p-values - which can be seen in Table V - derived from the paired comparisons indicate an extremely low likelihood that the observed differences occurred by chance, consistently rejecting the null hypothesis that the Ensemble classifier does not outperform the base models.

4. CONCLUSION

We presented a full pipeline, from web scraping the training *corpus* to the evaluation of distinct machine learning models capable of classifying Brazilian surnames by ethnic origin. This study explores the data processing steps in order to contribute to the transparency, reproducibility, and potential refinement of similar approaches, particularly in Brazilian and Portuguese-speaking contexts.

We hypothesized that LIGA would outperform the more traditional n-gram frequency classifiers on

Table V. Paired one-tailed t-tests (holdout macro F1-score; ensemble as reference)

Compared Model	t-statistic	p-value	**Significant ($p < .05$)**
`NGram (rank)`	100.5770	9.57×10^{-39}	✓
`NGram (cosine_tfidf)`	144.8867	2.47×10^{-43}	✓
`NGram (match_rate)`	200.3153	2.07×10^{-47}	✓
`LIGA (plain)`	281.8118	1.05×10^{-51}	✓
`LIGA (median)`	121.0942	4.45×10^{-41}	✓
`LIGA (log)`	289.8305	4.64×10^{-52}	✓
`LIGA (log+median)`	260.7748	9.92×10^{-51}	✓

surnames - as they seemed to fit the *short and noisy text* descriptor for which LIGA was designed - but the experiments proved them to be comparable to worse for this application. Our ensemble method outperformed all simple models evaluated, which shows that it is possible to achieve statistically significant improvements by combining different n-gram-based models. However, the practical gains in performance observed are somewhat modest in relation to the increased computation costs.

As future work, we propose more investigation into optimizing ensemble learning based on n-grammatical classifiers, both for performance and use of resources, including employing feature selection and hyperparameter tuning methods. Given the restricted amount of information that isolated surnames carry, we suggest that incorporating phonetic representations of graphemes as features (e.g., *Soundex*, *Metaphone*) and modern vectorization techniques may also yield relevant results.

REFERENCES

Cavnar, W. B. and Trenkle, J. M. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. pp. 161–175, 1994.

CPDOC. Dicionário histórico-biográfico da primeira república. imigração, 2000. Available at: `https://cpdoc.fgv.br/sites/default/files/verbetes/primeira-republica/IMIGRA%C3%87%C3%830.pdf`. Accessed on June 4, 2025.

Heringer, R. Affirmative action policies in higher education in brazil: outcomes and future challenges. *Social Sciences* 13 (3): 132, 2024.

IBGE. *Brasil: 500 anos de povoamento*. IBGE, 2007.

IBGE. *Características étnico-raciais da população : um estudo das categorias de classificação de cor ou raça : 2008*. IBGE, 2011.

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

Monasterio, L. M. Sobrenomes e ancestralidade no brasil. Tech. rep., Instituto de Pesquisa Econômica Aplicada (Ipea), 2016.

Nelson, J. R. and Shekaramiz, M. Authorship verification via linear correlation methods of n-gram and syntax metrics. In *2022 Intermountain Engineering, Technology and Computing (IETC)*. IEEE, pp. 1–6, 2022.

Ribeiro, C. A. C. and Carvalhaes, F. Research on social stratification in brazil. *Sociology Compass* 18 (9): e13266, 2024.

Schwartzmann, S. Fora de foco: diversidade e identidades étnicas no brasil. *Novos Estudos CEBRAP* vol. 55, pp. 83–96, 1999.

Tromp, E. and Pechenizkiy, M. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*. sn, pp. 27–34, 2011.

Vogel, J. and Tresner-Kirsch, D. Robust language identification in short, noisy texts: Improvements to liga. In *Proceedings of the 3rd international Workshop on Mining Ubiquitous and Social Environments*. pp. 43–50, 2012.