Classification of Endangered Languages Using Decision Tree Based Algorithms

Dongsheng Che
Department of Computer Science
East Stroudsburg University
East Stroudsburg, PA 18301

Taylor Shafer
Department of Computer Science
East Stroudsburg University
East Stroudsburg, PA 18301

Pu Tian
Department of Computer Science
East Stroudsburg University
East Stroudsburg, PA 18301

Abstract*— *Over all 7,000 languages around the world, nearly one third of them are at the risk of extinction while some others may suffer from other vulnerable dangers. Global efforts are needed to conserve the human language and culture. In this paper we selected the features with suggested causes of the danger of those languages from the Cambridge Handbook of Endangered Languages, and the actual data were obtained from various authoritative sources. We analyzed the data to check which features contributing the classification of endangered languages, we found that GDP, latitude and longitude were associated with language endangerment. Finally, we applied decision tree algorithm and its ensemble learning algorithms including Bagging, RandomForest and AdaBoosting for classification. We compared the performance of the four algorithms, and our evaluation indicates that ensemble algorithms overall have better higher accuracy rate compared to the base classifier, decision tree.

I. Introduction

According to the Cambridge Handbook of Endangered Languages, there are over 7,000 languages spoken in the world today. However, roughly one third of the worlds languages are in danger of extinction [1]. Entire language families are also at risk, as 15% of the world's linguistic stocks have become extinct and another 27% are now moribund [2]. Global efforts such as the Endangered Languages Project are underway to conserve human language and culture [3], but the scale of the problem makes complete coverage of at-risk languages impossible.

Primary causes of language endangerment as laid out in the Cambridge Handbook include natural disasters and disease, war and genocide, overt repression, and cultural, political, or economic dominance [1]. While well-known, these causes can be difficult to quantify, though efforts to do so are not unprecedented. The 2014 paper "Global distribution and drivers of language extinction risk", for instance, included an analysis of possible drivers of language extinction, including population size and economic growth [4]. Another recent study has made steps toward the modeling and quantitative analysis of language shift in particular, starting with a case study of Indonesian languages and social data [5].

While the study of languages continues to grow, overall availability of data has proven to be a challenging issue in language research. Databases of endangered languages necessarily rely on numerous outside studies for their information, some of which may be out of date. Linguists are often additionally limited to data collected for biased purposes, and a religious "commitment to missionary work" is apparent even in the Ethnologue: Languages of the World, one of the largest and oldest catalogues of language data [6]. This issue is currently being addressed in projects such as the Catalogue of Endangered Languages, which acknowledges levels of uncertainty in its own catalogue [7].

The aim of this study is to collect high-level data chosen based on the Cambridge Handbooks suggested causes of endangerment and availability of data, analyze the collected predictors, and use machine learning algorithms to classify language endangerment.

Machine learning algorithms have been successfully used in many classification problems. The availability of Big Data in the past years and the future will make machine learning algorithms ven more powerful. Decision tree is one of powerful learning algorithms and it has been widely applied due to its easy interpretation and fairly high prediction accuracy. Decision tree approaches have been used in astronomy [8], manufacturing and production [9], medicine [10] and bioinformatics [11]. In addition, decision tree based ensemble learning can improve classification accuracy [12].

In this study, we focus on decision tree and its ensemble learning algorithms for endangered language classification. Our research showed a high correlation between language endangerment and national development, while other features such as risk of natural disaster were not effective predictors when considered at a national level. Furthermore, our decision tree algorithms showed powerful classification of language endangerment when multiple predictors combined.

The remainder of this paper is organized as follows. Section 2 describes the data collection, feature analysis and machine learning algorithms. Section 3 shows the results of the feature selection and subsequent modeling. In Section 4, we will conclude the paper with a discussion of future work.

II. Materials and Methods

A. Data Collection

An initial dataset provided by the Guardian contained 2,722 examples of endangered languages, including the language's central coordinates, geographic spread, number of speakers, and degree of endangerment as determined by the UNESCO

978-1-5386-2165-3/17/$31.00 ©2017 IEEE

TABLE I: Summary of features collected in this study

Feature	Type	Description	Reference
Gross Domestic Product	Numerical	A countrys total economic output for the year	[13]
Latitude and Longitude	Numerical	Geographic coordinates of a languages approximate central location	[14]
S2 Cell	Numerical	The latitude and longitude mapped to a single value along a curve	[14]
Country range	Numerical	The number of countries in which a language is spoken	[15]
Global Peace Index	Numerical	A composite score of a countrys relative peacefulness	[16]
Human Freedom Index	Numerical	A composite score of a countrys personal and economic freedoms	[17]
World Risk Index	Numerical	A composite score of a countrys vulnerability to natural disasters	[18]
Greenbergs Diversity Index	Numerical	The probability of two compatriots speaking a different mother tongue	[19]
Family size	Numerical	The number of related languages in a language family	[19]

language endangerment scale [7]. The UNESCO scale classifies endangered languages in one of five categories: Vulnerable (or "unsafe"), Definitely Endangered, Severely Endangered, Critically Endangered, and Extinct. The classification itself is based on nine different factors, including intergenerational language transmission, literacy, and total number of speakers.

Total number of speakers is a major factor of a language's endangerment status, but as indicated by UNESCO's nine contributing factors, it is not the only factor and does not provide a clear split between the categories. Classification of languages is nuanced, as are the causes of endangerment. The number of speakers was not selected as a final feature in the model as it may be considered a result of endangerment rather than a cause.

One additional feature provided with the initial data set was latitude/longitude pairs indicating the central location of a language's speakers. Geographic coordinates were also considered holistically as level 25 S2 cells via the S2 Geometry library, which allowed mapping the latitude and longitude to single values along a Hilbert curve [14].

For the purposes of this model, language range was determined by the number of countries in which a particular language is spoken. Here, language range does not refer to physical distance, but rather to the spread of a language across political boundaries.

Additional data was collected from publicly available sources on the Internet. GDP, often associated with cultural hegemony and globalization, has been considered a prime factor behind language endangerment, as confirmed in the 2014 paper Global distribution and drivers of language extinction risk [4]. GDP data from 2015 was collected from the World Bank's data catalog [13]. An average GDP value was calculated for languages which span multiple countries, and data for GDP was only available at the national level.

Overt repression was estimated using the Human Freedom Index, an index introduced by the Cato Institute in 2015 that considers "79 distinct indicators of personal and economic freedom" [6]. As with GDP, an average value was taken for languages which spread across international borders. Similarly, a country's propensity toward war was collected from the Global Peace Index, introduced in 2007 by the Institute for Economics and Peace [16]. Natural disasters were accounted for using the World Risk Index, which has been produced by the University of Stuttgart since 2011 and which scores countries based on their exposure, susceptibility, coping capacity, and adaptive capacity in regard to natural disasters [18].

Two measures of linguistic diversity were also considered, both collected from the Ethnologue [19]. The family size of endangered languages was collected to account for language isolation, and linguistic heterogeneity at national levels was considered through the Greenberg Diversity Index. Greenberg's language diversity index measures the probability of two randomly selected people in the same country speaking the same mother tongue.

Table I is the summary of the list of ten features used in our study. We dropping samples with incomplete data, and the final data set contains 2,227 examples.

B. Feature Selection

Features used in for learning models can be predictive, interacting, redundant or irrelevant, thus it is important to do feature importance analysis and feature selection. Feature selection has been successfully applied in strategic decision making [20], Multiclass Mahalanobis-Taguchi system [21], sentiment classification [22], and bioinformatics [23],

We can roughly group feature selection into three groups: filter, wrapper, and embedded. In the filter methods, the subset selection procedure does not depend on the learning algorithm. For example, information gain, Chi-square test and fisher score belongs to filter methods. This kind of method is fast. In contrast, the subset selection procedure in filter methods is based on classification performance. Examples of wrapper approaches include recursive feature elimination, sequential feature selection algorithms, genetic algorithms [24], [25].

Scikit-learn [26] Python package provides all three kinds of implementations. In this study, we chose the fastest and most simplistic selection method from Scikit-learn: univariate filtering in Scikit-learn. In an univariate filtering, an ANOVA was performed for each feature, and then ranked according to their F-statistic. The rankings were be used to choose the best k features accordingly.

C. Learning Algorithms

In this study, we used the decision tree and decision tree based ensemble learning algorithms for classification. We used

Scikit-learn [26] Python implementations for these multiclass classification algorithms. Each of algorithms is described below.

1) Decision tree: For any decision tree, non-leaf nodes represent features, leaf nodes represent class labels, and branch represent the levels of the feature nodes [27]. The decision tree algorithm implements a top-down greedy search schema to search through all possible tree spaces. It starts with all training set, and chooses the best feature as the root node. The features are evaluated based on information gain, which can be evaluated based on entropy or Gini index. In this study, we used the Gini index in the Python Scikit-Learn library. The Gini index is to measure impurity of dataset, which is defined as

$$G(t, D) = 1 - \sum_{l \in L(t)} (Prob(t = l))^2 \tag{1}$$

where t is target feature in the dataset of D, $L(t)$ denotes the set of target values. In our study, there are five possible target values, *i.e.*, vulnerable, definitely endangered, severely endangered, critically endangered, and extinct. $Prob(t = l)$ denotes the probability for the target $t = l$.

Finally, information gain is defined as the difference between the Gini index for the whole dataset, and the remainder, as defined as follows:

$$IG(d, D) = G(t, D) - rem(d, D) \tag{2}$$

$$rem(d, D) = \sum_{l \in L(d)} \frac{|D_{d=l}|}{|D|} \times G(t, D_{d=l}) \tag{3}$$

Essentially, remainder an aggregated Gini index for a particular feature that partitions the dataset.

The decision tree algorithm splits the whole data based on the possible values of the selected best feature (*i.e.*, highest information gain). If the all instances in a split data have the same target value, then the process stops in that branch, and that node is a leaf node. On the other hand, if the subset does contain instances at least two target values, then the splitting procedure continues.

2) Bagging: Bagging is an ensemble learning algorithm that takes the majority votes of multiple base classifiers [28]. In this study, we use decision tree as the base classifier. Each base classifier model is trained on a subset of the observed dataset D. The training set of each classifier model can be sampled by bootstrap sampling, *i.e.*, randomly selecting a subset of given dataset with replacement. Let's denote the classification for each base model as $H_j(x)$, which returns the one of five classes in our study. The final bagged classification model can be expressed as follows:

$$H(x) = \arg\max_{l \in L(t)} \sum_{j=1}^{K} \delta(H_j(x), l) \tag{4}$$

where $\delta(H_j(x), l)$ is the Kranecker delta function that returns 1 if the two parameters are same, or return 0 if they are not. K is the number of base classifiers, and in our study we used 20 base classifiers.

3) Random forest: Random forest is the decision tree based ensemble model [29]. Like Bagging, random forest uses bootstrap sample. In addition, random forest also randomly sample features. Typically, $\sqrt{N}$ features out of N features are randomly sampled for each base classifier. In this study, we trained 20 base trees for our random forest model. Like bagging, the final random forest model will take the majority of votes from each of base classifier.

4) AdaBoosting: AdaBoosting is an ansemble learning algorithm that pays more attention to previously mis-classified instances [30]. The algorithm starts by building the first base classifier on randomly sampled dataset (*i.e.*, equal weights on all instances). The subsequent classifiers will assign higher weights to misclassified instances, and thus will have high probability of being selected. The final classification model is based on each of base classifiers, with their weights denoted as follows.

$$\alpha_j = \frac{1}{2} \times ln(\frac{1 - \epsilon}{\epsilon}) \tag{5}$$

where ϵ is error rate, thus high error rate model will have low α_j, small contribution to the final model. The final model can be expressed as follows:

$$H(x) = \arg\max_{l \in L(t)} \sum_{j=1}^{k} \alpha_j \times \delta(H_j(x), l) \tag{6}$$

In this study, we trained 20 base trees for our adaboosting model.

III. Results

A. Global map of endangered languages

Out of 2,227 endangered language, 549 are vulnerable, 571 are definitely endangered, 426 are severely endangered, 473 are critically endangered, and 208 are extinct. We mapped all these endangered languages as shown in Figure 1. As we can see from the map, there exist clusters of endangered languages across the globe, such as the cluster of extinct languages along the western coast of the United States and a dense group of endangered languages along the Himalayas.

B. Feature selection

We used univariate filtering to rank all 10 features used in this study. The feature importance rankings were shown in Figure 2. In agreement with previous research [13], GDP was overwhelmingly the best predictor of endangerment status with a score of 20.90. The second highest ranking feature was human freedom index. Latitude and longitude values also scored sufficiently well, with latitude scoring higher, while more specific locations as recorded in S2 cells scored lower than either. Purely linguistic features such as linguistic diversity and language isolation scored poorly. Despite being a notable factor in the endangerment of several languages in specific instances, overall disaster risk proved to be a poor indicator as well.

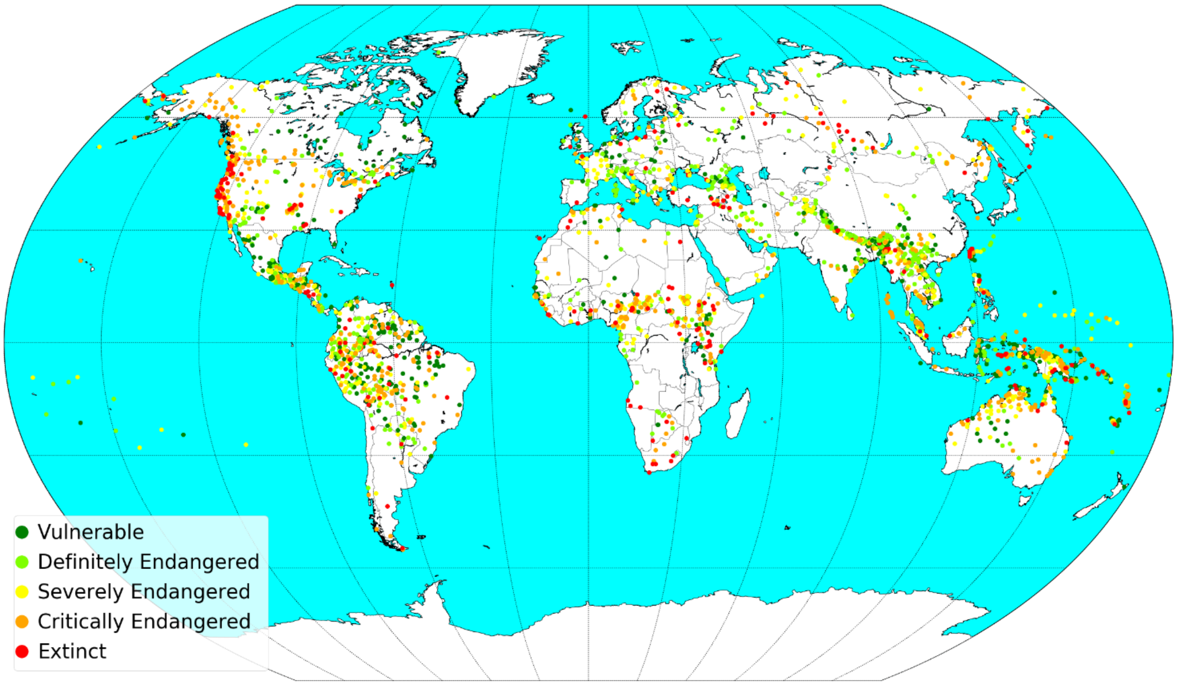


Fig. 1: Map of endangered languages by endangerment status

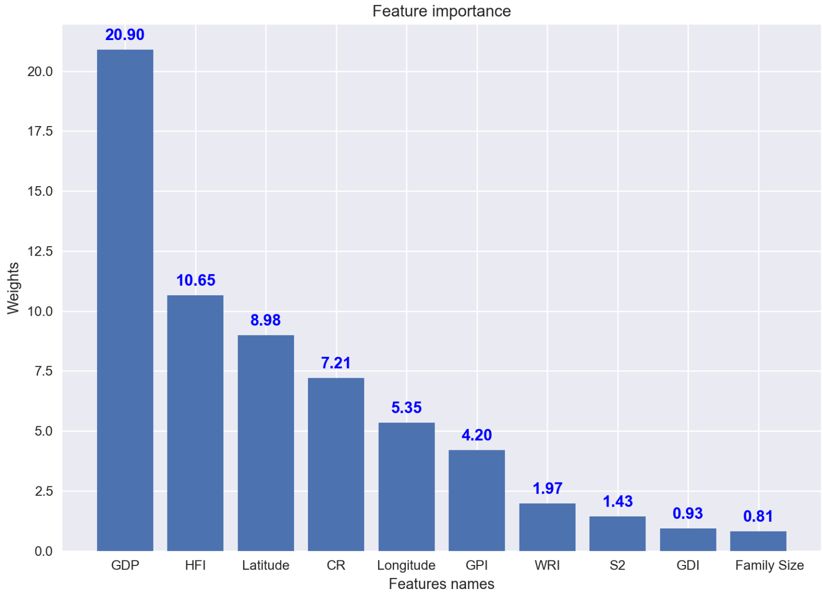
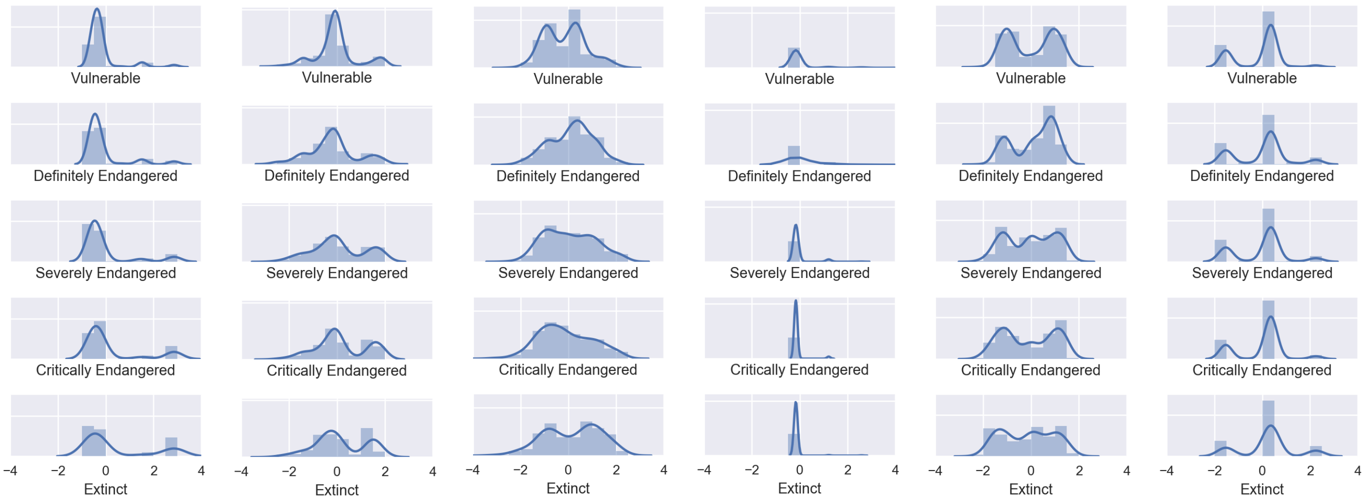


Fig. 2: Features importance ranking using univariate filtering



(a) GDP (b) HFI (c) Latitude (d) CR (e) Longitude (f) GPI

Fig. 3: Feature value distributions in each of five endangered languages

In addition to feature ranking based on scores, we also plotted the top-ranked feature value distributions in five different classes of endangered languages. As we can see from Figure 3, there are some distribution differences between severely endangered /critically endangered and the other three classes when considering the feature of latitude. However, overall there were highly overlapped regions of five classes of endangered languages for all feature used in this study, even for the highest ranking feature, GDP. This indicates that no single features are effective enough to classify the endangered languages, and the importance of combining multiple features. Given that, we chose the best six features for use in our classification algorithms, including GDP, Human Freedom Index, latitude, country range, longitude, and Global Peace Index.

C. Classification of endangered languages

To improve the classification accuracy, we used various machine learning algorithms that combine multiple predictors, and measured the algorithm performance based on the metrics of $recall$, $precision$, $F1$ score, $accuracy$, ROC curve and confusion matrix.

The confusion matrix is a table that is used to display the actual and predicted point in the relevant grid. Each cell in confusion matrix, denoted as c_{ij}, represents the number of actual class i that is predicted to be class j. Thus, the cell in the diagonal, denoted as c_{ii}, represents the number of correctly predicted instance for class i, where the remaining cells are the numbers of incorrectly predicted instances. Given that, $recall$, $precision$, $F1$ score, $accuracy$ are defined as follows:

$$recall = \frac{c_{ii}}{\sum_j c_{ij}} \tag{7}$$

$$precision = \frac{c_{ii}}{\sum_i c_{ij}} \tag{8}$$

$$F1 = \frac{precition \times recall}{precition + recall} \tag{9}$$

$$accuracy = \frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}} \tag{10}$$

Initially, we applied Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), Linear Discriminate Analysis (LDA), Quadratic Analysis (QDA) and Decision Tree on our dataset, we found that decision tree had best performance (performance results not shown here). Figure 4 shows a graphic view of a decision tree on classification of endangered languages. Based on our initial classification results, we decided to use decision tree as a base line, and also tested three ensemble learning algorithms: AdaBoosting, Random Forest and Bagging, aiming to improve classification accuracy.

We used randomly extracted 80-20 training-testing split across the dataset, and ran 50 times, and then averageed $recall$, $precision$, $F1$ score, and $accuracy$ for these four algorithms. Table II shows the classification results for these four algorithms. From the results, it is shown that the ensemble algorithms (AdaBoosting, Bagging and Random Forest) have 2% to 4% accuracy improvement over the decision tree method.

We also used the area under the ROC (Receiver Operating Characteristic) curve (AUC) to measure the classification performance. Theoretically, a well-performing classifier should have a higher AUC value. For multiclass classification we measure AUC for each class, and then aggregate them using macro-averaging and Micro-averaging metrics. From Figure 5, we can also see the ensemble algorithms show better performance over the decision tree.

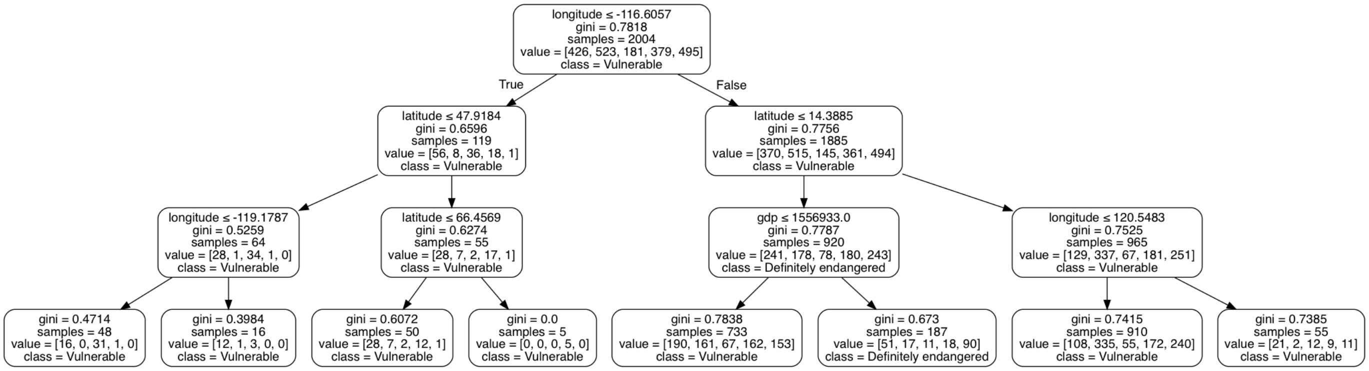
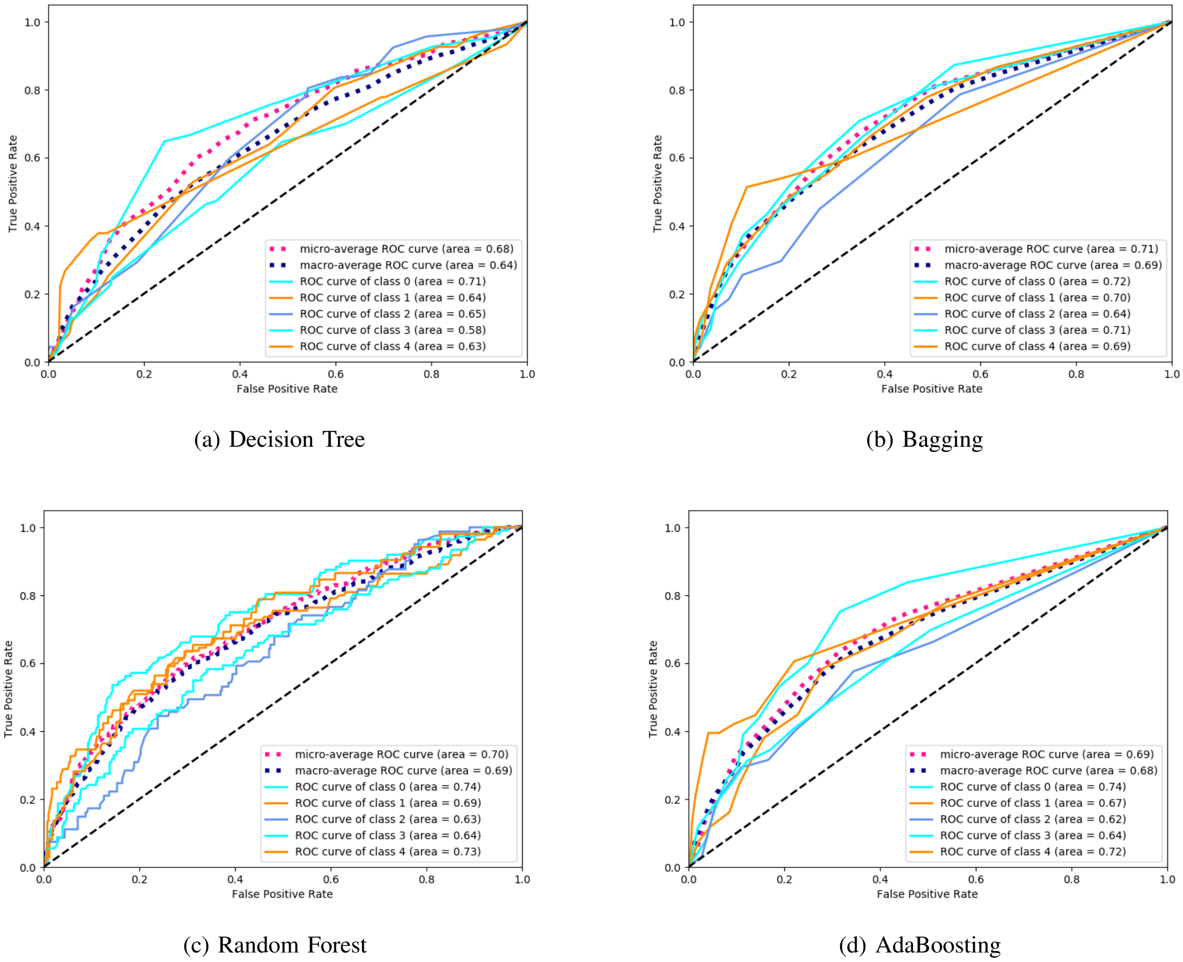


Fig. 4: Decision tree classification of endangered languages (depth=3)



(a) Decision Tree

(b) Bagging

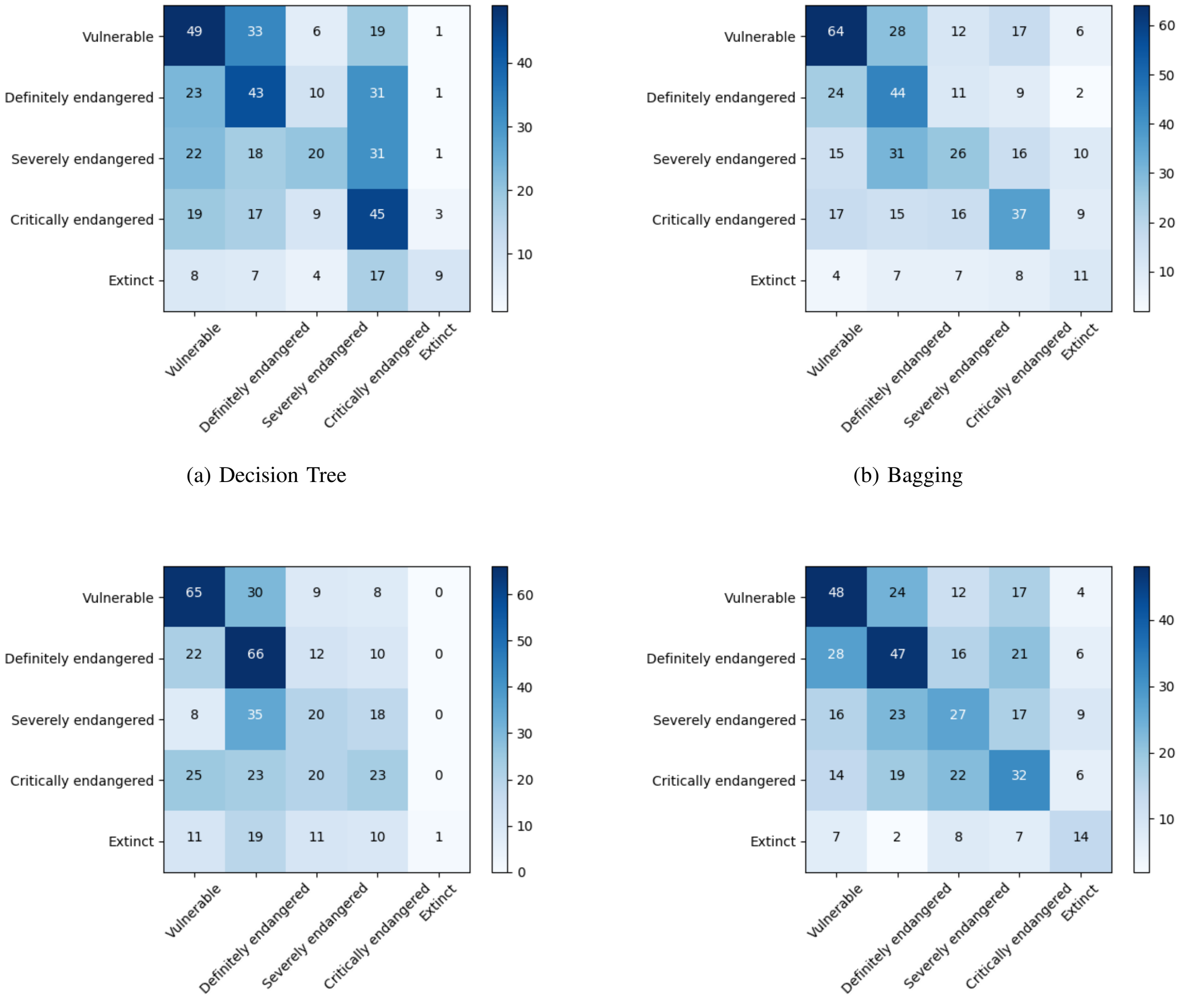
(c) Random Forest

(d) AdaBoosting

Fig. 5: ROC analysis using decision tree and its ensemble learning algorithms

TABLE II: Evaluation metrics using decision tree and its ensemble learning algorithms

Algorithm	**Recall**	**Precision**	**F1_Score**	**Accuracy**
Decision Tree	0.35	0.38	0.30	0.35
Bagging	0.39	0.38	0.37	0.39
Random Forest	0.37	0.39	0.32	0.37
AdaBoosting	0.39	0.38	0.37	0.39



(a) Decision Tree

(b) Bagging

(c) Random Forest

(d) AdaBoosting

Fig. 6: Multiclass confusion matrices using decision tree and its ensemble learning algorithms

Lastly, we used constructed the confusion matrices for each of the algorithms, so that we have better visualization for the performance of each algorithm. The values on the diagonal indicates the counts of correct classification. The higher of the value, the better of the model. From Figure 6, Bagging and Adboosting have higher diagonal values when compared to the values in the same rows, explaining why they have higher accuracy and ROC scores. Random performed slightly worse than these two ensemble algorithms, due to lower classification for severely endangered and extinct languages. Nevertheless, random forest also performs better than decision tree, evidenced by poor classification of severely endangered and extinct languages, and also not high number of mis-classification on vulnerable and definitely endangered languages.

IV. Conclusions and Discussion

In this paper, we collected and analyzed language endangerment associated dataset that suggested in previous studies, we have also used machine learning algorithms to classify endangered languages. Our experiments shown that decision tree was the best classifier. Our further experiments showed ensemble learning algorithms could improve classification accuracy by 2-4%. Overall, however, our model accuracy was still relatively low, and thus leaving the room for improvement through introducing more effective predictors.

The model's construction was additionally limited by the data set's composition. Only data from endangered languages was considered, so the model is not valid for examining languages which have not been previously classified as vulnerable or worse. Further, it is difficult to maintain up-to-date data on the various languages. Sources of aggregated language data such as the Ethnologue necessarily rely on individual language studies, which studies can take years to complete and may rely on estimates depending on the spread of the language. The status of a language can also change rapidly due to the sudden occurrence of natural disasters or disease. Though the World Risk Index was not useful for this model, it has been established that natural disasters do put endangered languages at great risk. For example, the 1998 earthquake in Papua New Guinea destroyed four villages of unique ethnic populations, leaving the fate of their languages uncertain even now [28].

As such, future studies could benefit from more specific data points, such as considering economic prosperity on a local or regional scale in addition to the GDP of entire countries. Historical data would also be useful in analyzing endangerment trends, as features like GDP, relative freedom, and relative peace can vary dramatically after significant events such as the fall of the Soviet Union.

The difficulty remains that such historical and specific data often does not exist in a usable, quantitative form. The World Risk Index and Global Peace Index used in this model were first launched within the past decade, and further analysis would be required to adapt the index to historical data. Therefore, further attempts to model language risk would benefit from additional studies in related fields, such as historical trends in economics and risk of natural disasters.

Lastly, state-of-the-art models such as recurrent neural network language model based phonotactic models that have been successfully used for spoken language identification [32], open another direction for endangered language classification.

References

[1] P. Austin and J. Sallabank, "The Cambridge handbook of endangered language", 1st ed. New York: *Cambridge University Press*, 2011, pp. 1-6.

[2] D. Whalen and G. Simons, "Endangered language families", *Language*, vol. 88, No. 1, , 2012, pp. 155-173.

[3] N. Lee and J. Van Way, "Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI)", *Language in Society*, vol. 45, no. 2, 2016, pp. 271-292.

[4] T. Amano, B. Sandel, H. Eager, E. Bulteau, J. Svenning, B. Dalsgaard, C. Rahbek, R. Davies and W. Sutherland,"Global distribution and drivers of language extinction risk, *Proceedings of the Royal Society B: Biological Sciences*, 2014, vol. 281, no. 1793.

[5] M. Ravindranath Abtahian, A. Cohn and T. Pepinsky,"Modeling social factors in language shift", *International Journal of the Sociology of Language*, 2016, vol. 2016, no. 242.

[6] "Human Freedom Index", The Cato Institute, 2017. [Online]. Available: https://www.cato.org/human-freedom-index.

[7] "Extinct Languages", Kaggle.com, 2016. [Online]. Available: https://www.kaggle.com/the-guardian/extinct-languages.

[8] N. Weir, U. M. Fayyad and S. Djorgovski."Automated star/galaxy classification for digitized POSS-II." The Astronomical Journal, 1995, 109(6):2401.

[9] S.K. Das and S. Bhambri."A decision tree approach for selecting between demand based, reorder and JIT/kanban methods for material procurement." Production Planning and Control, 1994, 5(4):342.

[10] I. Kononenko.,"Inductive and bayesian learning in medical diagnosis.", *Applied Artificial Intelligence*, 1993, 7(4):317–337.

[11] D. Che, C. Hockenbury, R. Marmelstein, and K. Rasheed. "Classification of genomic islands using decision trees and their ensemble algorithms", *BMC Genomics*, 2011, 11(Suppl 2):S1.

[12] D. Che, Q. Liu, K. Rasheed and X. Tao. "Decision tree and ensemble learning algorithms with their applications in bioinformatics", in *Software Tools and Algorithms for Biological Systems*, Springer Publisher, 191-199.

[13] "GDP Ranking", The World Bank, 2017. [Online]. Available: http://data.worldbank.org/data-catalog/GDP-ranking-table.

[14] O. Procopiuc, "Geometry on the Sphere: Google's S2 Library", 2011.

[15] T. Amano, B. Sandel, H. Eager, E. Bulteau, J. Svenning, B. Dalsgaard, C. Rahbek, R. Davies and W. Sutherland, "Global distribution and drivers of language extinction risk", *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, 2014, no. 1793.

[16] "Global Peace Index", Vision of Humanity, 2017. [Online]. Available: http://static.visionofhumanity.org/page/indexes/global-peace-index.

[17] "Human Freedom Index", The Cato Institute, 2017. [Online]. Available: https://www.cato.org/human-freedom-index.

[18] "World Risk Index", University of Stuttgart, 2017. [Online]. Available: http://www.uni-stuttgart.de/ireus/Internationales/WorldRiskIndex/.

[19] "Ethnologue: Languages of the World, *Ethnologue*, 2017. [Online]. Available: https://www.ethnologue.com.

[20] D. Y. Eroglu, K. Kilic, "A novel Hybrid Genetic Local Search Algorithm for feature selection and weighting with an application in strategic decision making in innovation management", *Information Sciences*, vol.405, pp.18-32, 2017.

[21] C. Su and Y. Hsiao, "Multiclass MTS for Simultaneous Feature Selection and Classification," *IEEE Trans. Knowledge and Data Engineering*, Vol. 21, pp.192-205, 2009.

[22] Y. Liu, J.-W. Bi, Zh.-P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms", *Expert Systems with Applications*, vol.80, pp.323-339, 2017.

[23] L.P. Wang, Y.L. Wang, C. Qing, "Feature selection methods for big data bioinformatics: a survey from the search perspective", Methods, vol.111, pp.21-31, 2016.

[24] Oh, I.S. Lee, J.-S., Moon, B. R, "Hybrid genetic algorithms for feature selection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, pp.1424-1437, 2004.

[25] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu, G. Forman, "Evolving feature selection", *IEEE Intelligent Systems*, vol.20, pp.64-76, 2005.

[26] "scikit-learn:Machine Learning in Python", 2017. [Online]. Available:http://scikit-learn.org/stable/

[27] J.R. Quinlan,"Induction of decision trees", *Machine Learning*, 1, 1986, 81-106.

[28] L. Breiman,"Bagging Predictors", *Machine Learning*, 24, 1996, 123-140.

[29] L. Breiman,"Random Forests", *Machine Learning*, 2001, 45, 5-32.

[30] Y. Freund, and Schapire, R. , "A decision-theoretic generalization of on-line learning and an application to boosting.", *European Conference on Computational Learning Theory*, 1995, 23-37.

[31] D. Crystal, "Language death", 1st ed. Cambridge [u.a.]: Cambridge Univ. Press, 2010.

[32] B. Srivastava, H. Vydana, A. Kumar Vuppala,M. Shrivastava, "Significance of neural phonotactic models for large-scale spoken language identification", *IJCNN*, Alaska USA, May 2017