Methods of Detecting Home Language Shift in Canadian Census Data

Chris Choy, Kiefer Co, Matthew Fogel, Clarke Garrioch and Katie Martchenko
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
Email: kleung@cs.umanitoba.ca

Abstract—**Canada is a nation composed of a highly diverse language population. This provides a unique opportunity to study the factors causing certain languages and language families to be lost over subsequent generations amongst allophones (people with a mother tongue other than English or French). This paper applies and compares the performance of Decision Tree Induction, Random Forest, and Categorical Naive Bayesian algorithms to census microdata to analyze the influence of various social and economic factors on the probability that allophones adopt official languages as their language spoken at home.**

Keywords-**component; language cohorts; allophones; mother tongue; language persistence**

I. INTRODUCTION

Canada is a nation composed of a highly diverse language population. Immigration and migration has risen both in numbers and as culturally relevant components of modern communities, especially in diverse countries such as Canada.

This provides a unique opportunity to study the rates at which certain languages and language families are lost over subsequent generations as allophones (people with a mother tongue other than English or French) adopt English or French as their primary language. Certain factors such as sex, age, educational status and economic success may prove to be key indicators of how quickly an individual adopts a language other than their mother tongue in everyday life.

A language shift occurs when an allophone adopts an official language as their primary language, i.e. language spoken at home. Several studies have aimed to measure language shift rates through linear regression on various cohorts of the population. Ultimately, it is impossible to ascertain precisely when a language shift occurs, so the insights offered by linear regression are limited in accuracy.

This paper proposes an application of various data mining algorithms and compares their accuracy and speed when used on census data, namely the Random Forest algorithm, the Decision Tree algorithm, and the Categorical Naive Bayesian algorithm.

Some challenges working with census microdata include the fact that the Public Use Microdata Files (PUMF) have been downsampled from the census population size. In order to perform an analysis that would take into account the population distribution, records are needed to be multiplied by a corresponding weight included in the PUMF datasets. Additionally, the PUMF datasets contain over 100 potential features, some of which contain largely invalid/unavailable data. As a result, some level of feature selection is required.

The Canadian census is conducted every five years. As a result, changes that occur between census periods may not be captured at the exact time of their occurrence.

II. RELATED WORK

Over the past several years, researchers have come up with multiple approaches to analyze census microdata to determine the rates at which allophones express a language shift. Several authors has focused on identifying whether a shift towards official languages (English and French) have occurred by determining if the mother tongue is the same as the language spoken at home. This is typically accomplished by performing linear regression.

A. Linear Regression of Language Cohorts

1) Fictitious Cohorts: Patrick Sabourin and Alain Bélanger use the concept of a 'fictitious cohort', in which groups separated by age or time since immigration are compared across a single census. They define language persistence as the proportion of each cohort that has kept their mother tongue as the language most often spoken at home. The authors analyze language shift using linear regression and polynomial regression. They then construct a survival curve which determines the probability that each subsection of a cohort (such as a specific age group) will undergo a language shift. They make the assumption that the rate of language shift is constant among several censuses. [1]

One limitation to this method is that members of a cohort are defined in binary terms as having lost a language if they no longer speak it at home, while this process might occur gradually in real time. Additionally, certain language groups may experience different rates of language shift. [1]

2) Synthetic Cohorts: Marie T Mora, Daniel J Villa and Alberto Davila use an alternative method known as 'synthetic cohort' analysis on census data in the United States. Their paper aims to better understand the recent dynamic of language loss and intergenerational maintenance of Spanish in the U.S., and compare it to other non-English languages. In other words, exploring the retention or loss

of Spanish and other non-English speakers in the U.S., particularly among foreign-born and U.S.-born children with immigrant parents. [2]

The technique for analysis, synthetic cohort analysis, is based on data drawn from 1980, 1990, and 2000 United States Censuses. It creates a temporal representation of a population, over ten-year intervals. The authors track the reported language use of individuals starting at ages 5-7 and ending at ages 15-17 across two United States census. This selection is due to children being able to speak at ages 5-7 and languages tending not to be lost after ages 15-17. [2]

This is in contrast to cross-sectional methods which use data from only one Census period to analyze language shifts. Combining this method with the synthetic cohort, the paper argues that the dynamic in language shift is better predicted, supported by what has been observed in the U.S. [2]

There are still some difficulties with this approach. For example, 1980 and 1990 samples could have emigrations before 1990 and 2000, and from this the "true" cohorts of the foreign-born may not be entirely reflected. [2]

3) Limitations of Linear Regression: As seen above, multiple approaches to analyzing language cohorts temporally run into limitations in how census data is collected. Sampling errors and poorly worded census questions make it difficult to capture whether emigration has occurred between censuses. Even within a census, it is tempting to view language shift within a cohort as binary when this process occurs gradually.

Populations are dynamic, and multiple categorical variables influence whether a language that is the mother tongue is spoken at home. A decision tree can reveal which categorical variables determine whether mother tongue is retained as the language spoken at home.

B. Decision Tree Performance in Other Areas

1) Decision Trees in Student Performance Prediction: Decision trees have also been compared and applied in other fields. Osmanbegovic and Suljic compared the performance of an implementation of a decision tree algorithm, Naive Bayes algorithm, and a Multilayer Perceptron algorithm in predicting student performance by the prediction accuracy, learning time, and error rate. [3]

Osmanbegovic and Suljic used 12 input features such as gender, GPA, and whether or not the student had scholarships, and outputted whether or not the student would pass or fail. The output could also be classified by letter grades, but due to the disparity in the amount of data for each class, was not used. [3]

They found that the Naive Bayes algorithm managed to outperform the C4.5 decision tree algorithm implementation, J48, in both prediction accuracy, and error rate. Additionally, it was found that the Naive Bayes algorithm and the decision tree algorithm created prediction models that were both accurate, and user-friendly enough for the stakeholders. [3]

2) Improvements: Similarly to Osmanbegovic and Suljic's data, some classes in census data are going to have differing amounts of data, and will affect prediction accuracy. Instead of changing the class groupings to get more equal representation, weights could be added to the training data to account for these differences. In the case of the Canadian Census data, these weights are already added to the data for use.

C. An Alternative Approach to Census Data Mining

In their paper, Klösgen and May took a different approach to mining census data. Klösgen and May used the United Kingdom's census data, which was only available in an aggregated form. The census data was available aggregated across various wards within the region, along with a detailed set of geographic layers. Thus, Klösgen and May decided to use those wards as the focus of their examination, and propose an application of SubgroupMiner, an advanced subgroup mining system. [4]

D. An Improved Decision Tree Algorithm

Hulten, Spencer, and Domingos' CVFDT algorithm improves on the VFDT decision tree learner by accounting for data changing over time. [5] Their CVFDT algorithm works by maintaining a decision tree with respect to a sliding window of data and grows an alternative subtree to replace an old one if it becomes out of date. This allows for it to learn a similar model to one from the VFDT algorithm but in constant time. [5]

Applying the algorithm to census data seems like a natural fit, and can be used to improve the performance of any decision tree learners in use. The addition of the time aspect could also be used to improve the accuracy in cases where the data from only one period of time is used, instead of multiple.

III. METHODOLOGY

A. General Process

The general process is as follows:

1) Determine when a shift in language occurs
2) Preprocess the data, converting any continuous values into discrete values
3) Split the data into training and testing sets
4) Construct the various classifiers, and evaluate them

Due to the nature of decision trees, all continuous values must be converted into a discrete value. Additionally, differences in datasets may result in the definition of language shift needing to be adjusted.

B. Mining the 2016 Canadian Census

In addition to the general steps stated above, applying them to the 2016 Canadian Census required further work, namely the additional preprocessing due to the format of the data.

The analysis was performed in Python 3 via Jupyter Notebooks. The **pandas** package was used to import the 2016 PUMF Individuals CSV file as a *DataFrame* and perform manipulations on the *DataFrame*, described in more detail in the data preprocessing section.

Several modules in the **scikit-learn** package were used for the classification algorithms in question, including the *DecisionTreeClassifier* for classifying based on decision trees, the *RandomForestClassifier* for classifying using random forests, and *CategoricalNB* for classifying using a Naive Bayesian algorithm. More details on these classifiers can be found in their respective sections.

The **bokeh** package was used to generate a bar chart describing the features with the greatest amount of invalid data and to visualize the ROC AUC scores for the various classifiers. The **tabulate** package was used to create a table listing feature importances in descending order of importance. The **graphviz** package was used to visualize decision trees.

1) Data Format: Data was collected from the 2016 Canadian census public use microdata file (PUMF), which contains around 930,421 (or 2.7% of the target population) individual, anonymized records, with 123 features. Additionally, there is an individual weight attached to each record, and 16 estimate weights for sampling variability.

While the PUMF did give individual records, some of the data was aggregated to preserve confidentiality (e.g. categories being combined together), and some records had some of their variables changed to 'Not Available' for similar reasons. Furthermore, only the largest of the census metropolitan areas and provinces were covered.

Since the PUMF data was only a sample of the target population, each record includes an individual weight to indicate how much of the target population that the record represents. In addition to the weights, 31 of the features are drawn from the universe of family, household, and dwelling universes, and the remaining 92 features are from the individual universe.

2) Data Preprocessing: Prior to the execution of any of the classifiers, the PUMF dataset is passed through a single data preprocessing pipeline. This pipeline achieves two broad objectives, namely pruning uninteresting data and selecting against features containing a high proportion of invalid data.

The feature selection process performed was a mix of manually intuited choices and insights from initial analysis using *scikit-learn*'s decision tree classifier on unprocessed data. Features which were pruned were both uninformative with regards to the goal of detecting home language shifts and possessed clearly understood reasons for their behaviour. For example, MTNEN (mother tongue being English) was extremely negatively correlated with the shift metric, as English is the most spoken language in Canada and an unlikely home language to shift away from. Conversely,
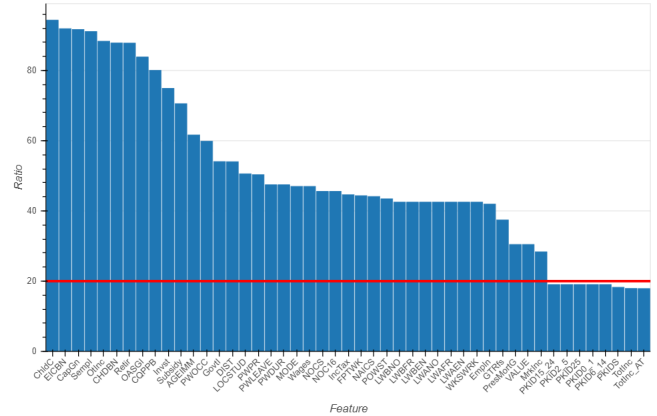

Figure 1. Invalid data by feature

LWAEN (language at work being English) and LWAFR (language at work being French), had little to no influence on the shift metric, which was understood to be the consequence of work language rarely being a language other than English or French. Such features were omitted from the preprocessed version of the dataset.

The initial data analysis also found promising improvements when dropping rows featuring English as a mother tongue. Removal of these rows lowered the unusable proportion of many features that was thought to provide useful insight, such as AGEIMM (age of immigration) which dropped from 88 to 61 percent unusable. Although AGEIMM in particular remained unsuitable for the analysis as only a fraction of respondents provided a usable value, a decision was made to remove English mother tongue rows from the dataset with the intention of providing more focused analysis on changes within Canada's non-official language communities, from which shifts to English as a home language were more common than the inverse.

The PUMF dataset also contained missing, invalid, or inapplicable values as specified by certain numeric codes (e.g. 88888888) as specified in its accompanying Study Documentation. The initial preprocessing involved detecting the number of such values in each feature column as a proportion of the total number of entries. Features which possessed an unusable percentage greater than 20% were pruned so as to not obscure the analysis by training classifiers on presence of a data feature rather than its content. This was done after the removal of English as a mother tongue rows described earlier. A subset of the feature's invalid data rates can be seen in figure 1.

Finally, the data is split into training and testing data sets with 75% of the data being allocated for training, and the remaining 25% for testing.

3) Imputation: Since some values in the PUMF data are marked as invalid or unavailable, further preprocessing on the data was required. As mentioned, data imputation was

considered as a remedy for unusable data, but opted against due to the presence of weights attached to each record in the PUMF dataset. The dataset is meant to balance presenting a representative subset of Canada's demographics while also anonymizing census results which, by Canadian law, are to be withheld from complete public disclosure for 92 years post-response. As such, each row in the PUMF dataset is a weighted representation of several individuals, listed under the attribute WEIGHT. Imputation of unusable values, even with the feature's weighted mean as a substitution, was reasoned to be inapplicable without rendering the dataset no longer representative of Canada's demographics.

4) Shift Detection: The process for identifying language shift involves constructing a Boolean feature called **languageShift**, computed from the MTNNO (mother tongue) and HLANO (home language) attributes.

In the PUMF dataset, both the MTNNO and HLANO features are categorical attributes encoded as a numeric value. Each feature's encodings are independent and cannot be readily compared, although there exists an intersection between the categories represented within each feature. Dictionaries for mapping MTNNO and HLANO categories to equivalency classes were manually defined using information outlined in the PUMF Study Documentation file. These classes include both whole languages such as "Arabic" as well as grouped categories such as "Austro-Asiatic languages".

As the MTNNO categories are less broad and appear to be a superset of the HLANO categories, some categories appear in MTNNO but not HLANO, such as "Uralic" or Korean". Rather than omitting rows containing these mother tongues, they are categorized under the "All other languages" class.

The value of the languageShift attribute is then defined as the presence of inequivalent class mappings in the MTNNO and HLANO features within a data row. The reasoning behind this decision is that a difference between mother tongue and home language implies a shift having occurred at some point within an individual's lifetime, with the assumption made being that an individual's mother tongue must at one point have been their home language, possibly only in childhood or even outside of Canada.

If MTNNO can be mapped to HLANO, the languageShift value of a row is false. If no mapping can be found, the languageShift value will be true. As with other features in the dataset, the results of the languageShift column are to be interpreted as representative of several individuals. A row's weight is to be factored into any analysis on or prediction of the languageShift feature. After this preprocessing, the counts of the true class (language shift has occurred) and the false class (language shift has not occurred) are 306,645 and 72,430, respectively.

C. Decision Tree

The decision tree classifier was build using *scikit-learn*'s implementation, *DecisionTreeClassifier*. To prevent overfitting of the training data, the maximum depth of the decision tree was set to 8 since this is approximately the square root of the number of features being included in the classifier.

The languageShift and WEIGHT attributes are first separated from the dataset to serve as class labels and weights respectively. MTNNO and HLANO are dropped from the dataset as they are directly involved in the computation of languageShift, for which their effects on a classifier for would be uninteresting. The remaining attributes in the preprocessed dataset are then to be used as classification features. The features and weights for the training data are then passed into the classifier, with the weights being passed into the *sample weight* parameter of the *fit* function on the decision tree classifier. The class of test samples are predicted using the *predict* function. Finally, the results of the classifier are output in a confusion matrix.

D. Random Forest

Similarly to the decision tree classifier, the random forest classifier was constructed using *scikit-learn*'s implementation, *RandomForestClassifier*, using 20 estimators and a controlled random state to maintain similar, repeatable results after multiple runs. As Random Forest classifiers inherently have measures to reduce overfitting of the training data, no maximum depth was specified.

The classifier is trained on the training data set, along with their weights, then evaluated on the testing data set. The predicted class of test samples is determined using the *predict* function by selecting the class with the highest mean probability estimate across all trees in the random forest. The results are also outputted to a confusion matrix.

E. Naive Bayesian

The Naive Bayesian classifier was built using *scikit-learn*'s *CategoricalNB* - a Naive Bayes classifier for categorical features. This classifier assumes that each feature has its own categorical distribution. The *fit* method was used to train the classifier on the weighted training data. The *predict* method was used to perform classification on the testing data. The results are also outputted to a confusion matrix.

IV. ANALYTIC EVALUATION

To compare the different classifiers, metrics such as precision, recall, F1 score, and AUC ROC scores were used. Additionally, the top 10 feature importances for the decision tree classifier, and the random forest classifier were examined to further compare the two. The metrics can be seen in figures 2, 5, and 8. Additionally, a graphical representation of the AUC ROC scores can be seen for each of the classifiers in figures 3, 6, and 9. Finally, graphs

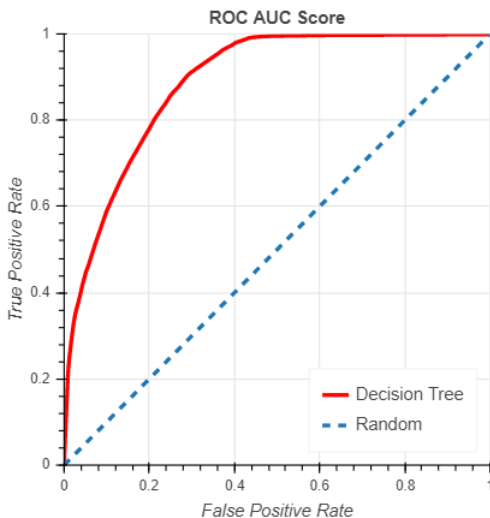|	precision	recall	f1-score	support
False	0.87	0.96	0.91	76719
True	0.70	0.41	0.52	18050
Total				94769
Accuracy	0.85			
ROC AUC Score	0.89			

Figure 2.   Decision tree metrics

Figure 3.   Decision tree ROC AUC curve

comparing the classifiers by each of the metrics can be found in figures 10, 11, 12, and 13.

As shown in figures 10, 11, 12, and 13, the random forest classifier only provided a slight improvement on the standard decision tree classifier. Additionally, the top 10 most important features between the two classifiers are very similar (see figures 4, 7).

Since the target class is highly unbalanced, with roughly 81% of the data belonging to the false class, accuracy is perhaps not the best metric to use to compare the different models. The other metrics calculated such as F1-scores and AUC ROC scores are better for our models. The ROC AUC scores for the Decision Tree model and Random Forest model are 0.88 and 0.89, respectively. This means they are performing similarly in determining the positive class from the negative class. The ROC AUC score for Naive Bayes is 0.82, 8 percent lower than the two tree-based models. This is a significant difference, and it is reflected in the ROC AUC curve, where a noticeable dip in the curve occurs early on when compared to the tree-based models' curves. From this we can say the Naive Bayes classifier performs significantly worse at determining the the positive class from the negative class than the tree-based models.

Feature	Importance
GENSTAT (Generation status)	0.459719
VisMin (Visible minority status)	0.12905
PR (Province, current)	0.0875603
POB (Place of birth)	0.0789063
ETHDER (Ethnic origin)	0.0784669
IMMCAT5 (Immigration Admission category)	0.0324108
LOC_ST_RES (Location of study)	0.0279763
DPGRSUM (Population group)	0.0245261
LFACT (Labour force status)	0.0120307
AGEGRP (Age)	0.0118114

Figure 4.   Decision tree feature importances

	precision	recall	f1-score	support
False	0.88	0.96	0.92	76719
True	0.72	0.46	0.56	18050
Total				94769
Accuracy	0.86			
ROC AUC Score	0.90			

Figure 5.   Random forest metrics
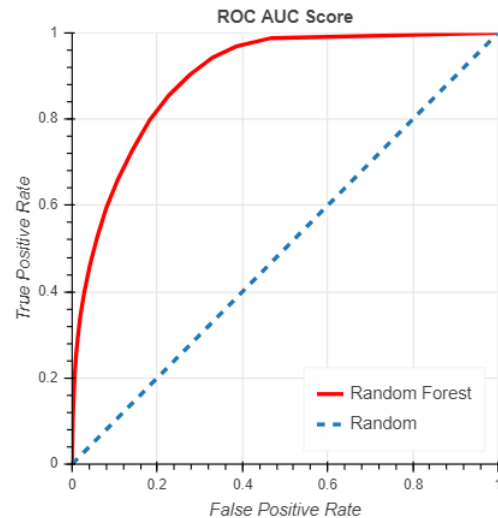
Figure 6.   Random forest ROC AUC curve

Feature	Importance
GENSTAT (Generation status)	0.062281
ETHDER (Ethnic origin)	0.0573086
PR (Province, current)	0.0523802
POBM (Mother's POB)	0.048009
POB (Place of birth)	0.0405777
POBF (Father's POB)	0.0321791
VisMin (Visible minority status)	0.0253287
AGEGRP (Age)	0.0240645
PR5 (Province, 5 years ago)	0.0239604
IMMCAT5 (Immigration admission category)	0.0232537

Figure 7.   Random forest feature importances

	precision	recall	f1-score	support
False	0.93	0.68	0.79	76719
True	0.37	0.80	0.51	18050
Total				94769
Accuracy	0.71			
ROC AUC Score	0.82			

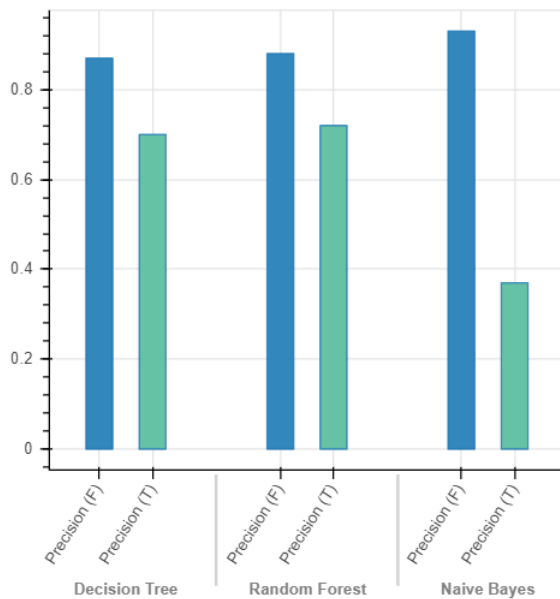Figure 8.   Naive Bayes metrics
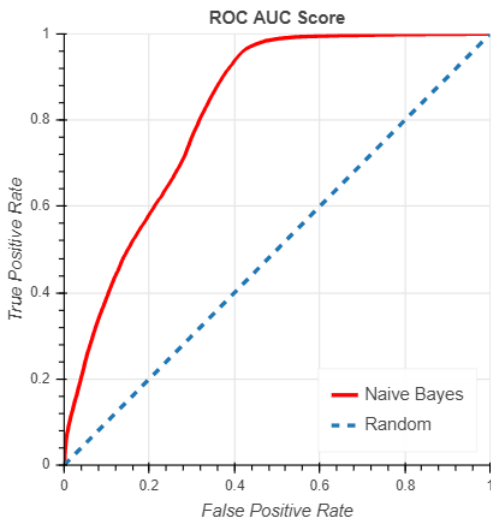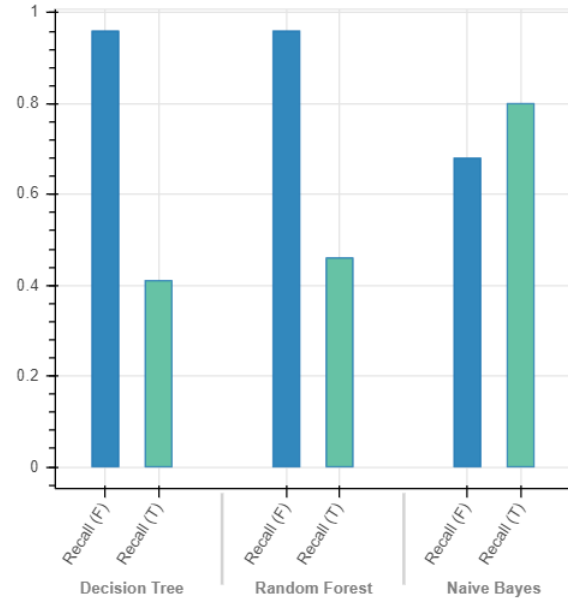

Figure 9.   Naive Bayes ROC AUC curve

Figure 10.   Classifier precision values

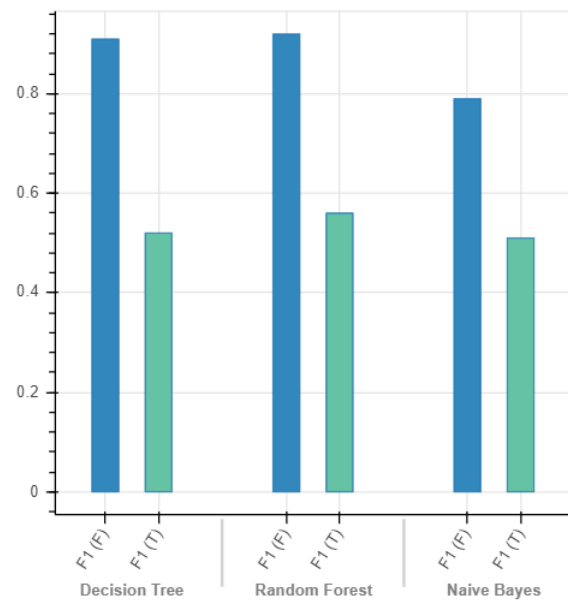Figure 11.   Classifier recall values

Figure 12.   Classifier F1 values

V. CONCLUSION

A. Future Work

One aspect that could be improved on in future work is to tweak the hyper parameters for the classifiers via cross validation. For instance, *scikit-learn*'s decision tree classifier implementation allows for setting a maximum depth on the resulting tree. Modifying this value may result in a more accurate classifier.

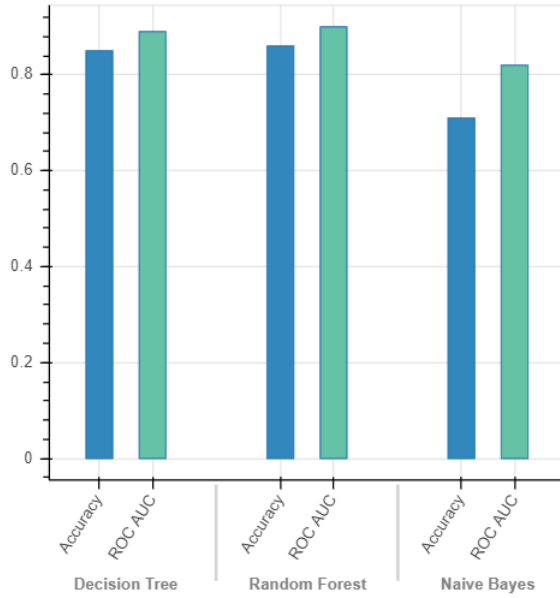Additionally, the 2016 Canadian Census data set contains

Figure 13. Classifier accuracy and ROC AUC values

REFERENCES

[1] P. Sabourin and A. Bélanger, "The dynamics of language shift in canada," *Population*, 2015.

[2] M. T. Mora, D. J. Villa, and A. Davila, "Language shift and maintenance among the children of immigrants in the u.s.: Evidence in the census for spanish speakers and other language minorities," *Spanish in Context*, 2006.

[3] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, 2012.

[4] W. Klösgen and M. May, "Census data mining - an application," 2012.

[5] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," 2001.

mostly data indicating a lack of language shift. As a result, the classifiers learn mostly on these negative results, and it may impact the rates for positive results. The weights on the training data could be re-weighted to account for this, but further complications may arise when trying to maintain the data being representative of the Canadian public.

Furthermore, this analysis could be applied over multiple census years. In addition to increasing the amount of data available, adding a temporal dimension could add an interesting aspect to the classifiers, possibly taking historical events into account.

Adding a geographical dimension to the classifiers could identify more granular language shifts in Canadian provinces and territories. It would be interesting to apply the analysis on municipalities such as the City of Winnipeg as well. The Winnipeg open census data portal contains information on language spoken broken down by neighbourhood and ward.

Another interesting metric that could be determined in future work is the conditional probability of each category of each feature influencing the language shift category. For instance, *scikit-learn*'s CategoricalNB classifier implementation contains a feature log prob attribute that returns a 2D array for each feature, which contains the logarithmic probability that each category of the feature influences the language shift category.

ACKNOWLEDGMENT