DETERMINING THE ETHNO-NATIONALITY USING WRITTEN ENGLISH TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Ethno-nationality is where nations are defined by a shared heritage, for instance it can be a membership of a common language, nationality, religion or an ethnic ancestry. The main goal of this research is to determine a person's country-of-origin using English text written in less controlled environments, employing Machine Learning (ML) and Natural Language Processing (NLP) techniques. The current literature mainly focuses on determining the native language of English writers and a minimal number of researches have been conducted in determining the country-of-origin of English writers.

Further, most experiments in the literature are mainly based on the TOEFL, ICLE datasets which were collected in more controlled environments (i.e., standard exam answers). Hence, most of the writers try to follow some guidelines and patterns of writing. Subsequently, the creativity, freedom of writing and the insights of writers could be hidden. Thus, we believe it hides the real nativism of the writers. Further, those corpora are not freely available as it involves a high cost of licenses. Thus, the main data corpus used for this research was the International Corpus of English (ICE corpus). Up to this point, none of the researchers have utilised the ICE corpus for the purpose of determining the writers' country-of-origin, even though there is a true potential.

For this research, an overall accuracy of 0.8140 for the flat classification (for all ten countries) and accuracy of 0.7549 1.000 for sub-categories were received. In addition, the best ML model obtained for the flat classification strategy is linear SVM with SGD optimizer trained with the character 9-gram model.

1 INTRODUCTION

Ethno-nationality can be determined as the identity in which jointly defined by ethnicity and nationality (Sapp, 2012). Identifying a writer's country-of-origin is a part of identifying his ethno-nationality (Jerry, 2008). Foremost focus of ethno-nationality in this research is to identify a writer's country-of-origin based on their writings in a non-native or second language. In recent years, identification of ethno-nationality of a writer has gained a growing interest. In author profiling demographic features (such as age, gender, education, native language, country of origin, etc.,) of an author from a written text will be identified, which are commonly needed in forensic linguistics. For instance, Intelligence to build a profile of their suspect, to identify the author of an anonymous email threat. Therefore, this will enable to limit the search space as well (Estival et al., 2007). Moreover, for business applications this can be useful; for example, in marketing where the demographic features as stated above of customers is important to predict behaviors, upgrade the current products and to develop new products.

People from different ethno-nationality make various language errors when learning a language. Identification of writer's country of origin could have an impact on educational applications designed towards non-native speakers of a language. Besides, this can be used as a plug-in to online tutor systems to provide more tailored feedback to the students about their mistakes (Tetreault et al., 2013). This can help researchers to identify specific teaching and learning issues in different ethno-nationalities. This will enable them to develop pedagogical learning materials to address and solve those issues.

In this research we focus on English as the second language. As English has become a 'universal communication' language due to the globalization. It is no longer restricted to the native countries such as England and United States (Eric, 2013). English language is now available in numerous emerging fields, and has become an essential requirement of labour market and further it is considered to have a cultural importance (Marko, 2009). We define, A controlled environment as a place or area where rules, regulations and norms are subject to strict enforcement. For instance, in an examination setting, candidates are required to produce answers limiting the scope of the question.

2 RELATED WORK

Distribution of English language around the world can be identified in two ways, mainly; based on the geographic distribution and based on the Kachru's theory (Braj B, 2004). In this study when segmenting according to the geographic distribution we have considered clusters namely; 'Asia', 'non-Asia', 'South Asia', 'non-South Asia' and 'North America'. According to the Kachru's 'Concentric Circles', English speakers were segmented in to three main categories, namely; 'inner circle', 'outer circle' and 'expanded circle'.

As stated by his model, the inner circle comprises of the countries where the English is the native language (i.e., UK, USA, Canada, New Zealand, etc.). The outer circle includes countries where there are small communities of native English speakers however English is used as the second language in education and official purposes (i.e., Sri Lanka, India, Singapore, Nigeria, etc.). The expanding circle contains countries which considered English as foreign language (i.e., China, Indonesia, Japan, Saudi Arabia, etc.).

The work of Koppel et al. (2005) is one of the first ethno-nationality identification work where it involves categorizing users based on their native language employing Support Vector Machines (SVMs) on various stylistic features focused on identification of common. Authors achieved 0.802 accuracy using five chosen languages (i.e., Czech, French, Bulgarian, Russian and Spanish) from International Corpus of Learner English (ICLE)(Granger, 2014). The ICLE dataset contains argumentative essays writings of the university students and less nativism and creativity of the writer involved.

Bykh & Meurers (2012) proposed use of recurring n-grams on three different classes (word based, POS based, Open-Class-POS-based) as features for training SVMs. Out of other SVM implementations LIBLINEAR produced best results. For this study, random data from seven native languages selected from the ICLE corpus. The highest performance was obtained for word-level n-grams with an accuracy of 0.8971.

Gebre et al. (2013) employs linear SVM, logistic regression and perceptron (as baseline) for the native language identification (NLI) and achieved accuracy of 0.814 for eleven languages of 'The Test of English as a Foreign Language' (TOEFL) data set(Blanchard et al., 2013). Features used includes; word n-grams, POS n-grams, character n-grams and spelling errors. TOEFL11 has become standard benchmark in NLI tasks since its introduction for the NLI Shared Task 2013. The main limitation of the TOEFL11 dataset is that it is collected in a more controlled environment (i.e., exam for English).

Cimino & Dell'Orletta (2017) utilizes a novel stacked classifier approach where linear logistic regression based sentence feature classifier is stacked with a SVM based document feature classifier with standard lexical, stylistic and syntactic features. However stacked classification approach has gained a minor gain compared to unstacked. Best results of the NLI shared task 2017 reported for this approach with F1-score of 0.8818 for the TOEFL11 dataset.

Kulmizev et al. (2017) introduced 'Groningen' system for the NLI Shared Task 2017 which out performs employing linear SVM for character 1-9 n-grams with the F1-score of 0.8756. Authors have reported that several experiments done with ensemble approach and other features such POS, word, lemma n-grams, skip-grams; and those failed to match the performance of character 1-9 n-gram system.

Goutte & Léger (2017) explored use of voting ensemble SVM models with character, word and POS n-grams. Authors confirms that ensemble methods provide minor but systematic predictive perfor-

Table 1: Summary of the current literature

Research	Corpus	Accuracy /F1-Score	#Classes	Approach
Koppel et al. (2005)	ICLE	0.802	5	Support Vector Machines (SVMs) on various stylistic features namely; function words, letter n-grams, and errors and idiosyncrasies
Bykh & Meurers (2012)	ICLE	0.8971	7	SVMs with recurring n-grams of three different classes (word based, POS based, Open-Class-POS-based) as features
Gebre et al. (2013)	TOEFL11	0.814	11	Linear SVM, logistic regressions and perceptron as the linear classifiers with word n-grams, POS n-grams, character n-grams and spelling errors
Cimino & Dell'Orletta (2017)	TOEFL11	0.8818 /0.8818	11	Two-stacked sentence- and document-feature based classifier architecture. Output of the sentence-level linear regression model is being used by a document-level SVM.
Kulmizev et al. (2017)	TOEFL11	0.8755 /0.8756	11	Linear SVM and character 1-9 n-grams
Goutte & Léger (2017)	TOEFL11	0.8736 /0.8740	11	Voting ensemble SVM models approach with character, word and POS n-grams features
Markov et al. (2018)	TOEFL11 and ICLE	0.4883 (TOEFL11) & 0.6948 (ICLE)	11 (TOEFL11) & 7 (ICLE)	SVM with one-vs-all (OvA) multiclass approach. Abstract POS n-gram and punctuation marks (PM) features have been used.
Malmasi & Dras (2018)	TOEFL11, EFCAM-DAT, ASK, JCLC	0.871 (TOEFL11)	11 (TOEFL11)	Supervised multi-class classification approach with feature including character, function word, POS n-grams, dependencies, CFG rules, adaptor grammars and TSG fragments used

mance gains. Highest F1-score performance was with best-vote approach consisting 10 models is 0.8740 for TOEFL11 dataset.

Markov et al. (2018) has used punctuation-based features with POS n-grams for his experiments and accuracies of 0.4883 (TOEFL11) and 0.6948 (ICLE) reported for best performing settings. SVM with one-vs-all (OvA) multi-class classification approach has been used to conduct these experiments.

Malmasi & Dras (2018) employed a supervised multi-class classification approach and incorporated several corpora, including; TOEFL11, EF Cambridge Open Language Database Corpus (EFCAM-DAT), ASK Corpus (Andrespråkskorpus, Second Language Corpus), and Jinan Chinese Learner Corpus. The features extracted were; word/lemma n-grams, character n-grams, function word n-grams, POS n-grams, dependencies, CFG rules, adaptor Grammars and TSG fragments. Highest accuracy of 0.871 is reported for the TOEFL11 dataset.

According to the literature it is evident that most of the researchers in the computational linguistic community have employed ICLE in the early stages and TOEFL in the latest NLI tasks (refer Table 1) which both were collected in more controlled environments. Finally, due to above controlled environments; being unable to capture real nativism in written texts certainly inflate the performance of ethno-nationality identification. On the other hand, identifying the nation of the English writer opposed to his native language will be equally beneficial for author profiling as well. Thus, we

introduce use of International Corpus of English (ICE) for the ethno-nationality identification. To the best of our knowledge, none of the researches have been conducted to determine the writer's country of origin based on the International Corpus of English (ICE) corpus. Moreover, very limited number of researchers have identified the significant features which helps to distinguish Sri Lankan English writers using a large corpus like ICE. One limitation of these researches raised by the authors Tofighı et al. (2012) is that, since for most of the web-based applications, automatic spell-checker has been applied, idiosyncratic features including misspelling and other anomalies are ignored. This may hide some of the features which will be useful in identifying the real categories. Nevertheless, in current context, usage of spell checkers and grammar checkers can be seen frequently. Thus, this will be a common limitation in similar types of researches.

3 RESEARCH METHODOLOGY

Supervised learning approach was chosen as it is suitable in classifying text documents into classes more accurately if the classes are known and the data set is labelled (Slotte, 2018). Hence, according to the literature most promising ML algorithms for the text classification such as; Support Vector Machines with Stochastic Gradient Descent Optimizer (linear SVM with SGD), Multinomial Naïve Bayes (MNB), Decision Tree (DT) and Random Forest (an ensemble approach) have been employed. For this research mostly Scikit-learn(Pedregosa et al., 2012) has been used for preprocessing, feature extraction and classification tasks. Scikit-learn is a Python module for machine learning which provides state-of-the-art implementations of many well-known machine learning algorithms, and maintains an easy-to-use interface(Pedregosa et al., 2012). The workflow of the research is depicted in Figure 1 (Left).

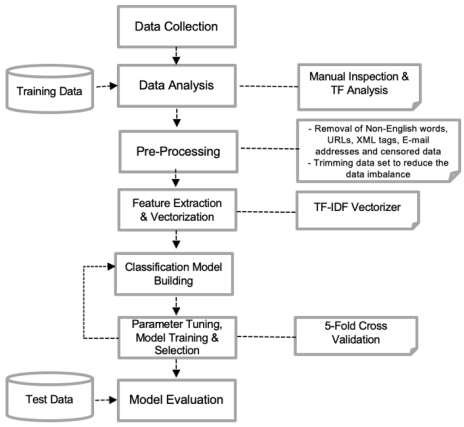
The main assumptions of this research are authors from the same ethno-nationality share the same linguistic features in their writing and will often have an influence on the way they express themselves in writings (Jain et al., 2017).

The research questions addressed in this research are;

Q1. How can texts produced by English writers in a given ethno-nationality be captured from existing corpora?

Q2. How do texts produced by Sri Lankan English writers differ from other English writers?

Q3. Which machine learning techniques can gainfully employ the extracted data to identify country-of-origin of English writers?



	Language	Original Total Doc & Word Count		Adjusted Total Doc & Word Count		Adjusted Training (Seen Data) 70%		Adjusted Testing (Unseen Data) 30%	
		#Docs	#Words	#Docs	#Words	#Docs	#Words	#Docs	#Words
1	Canada	200	439,535	1,158	439,461	810	305,836	348	133,625
2	Hong Kong	200	550,471	1,433	550,395	1,003	385,660	430	164,735
3	India	200	458,233	1,198	458,233	838	320,034	360	138,199
4	Ireland	200	426,113	1,112	426,113	778	297,332	334	128,781
5	Jamaica	200	422,465	1,104	422,465	772	296,822	332	125,643
6	Nigeria	510	400,597	1,159	400,597	811	279,145	348	121,452
7	Philippines	178	410,204	1,081	410,186	756	286,798	325	123,388
8	Singapore	200	441,383	1,155	441,383	808	309,950	347	131,433
9	Sri Lanka	199	473,946	1,233	473,913	863	331,777	370	142,136
10	USA	199	431,860	1,142	431,782	799	302,350	343	129,432
Total		**2,286**	**4,454,807**	**11,775**	**4,454,528**	**8,238**	**3,115,704**	**3537**	**1,338,824**

Figure 1: (Left) Workflow of the research methodology. (Right) Document & word distribution of ICE corpus
#Docs = number of documents, #Words = number of words

3.1 DATA COLLECTION

The main dataset employed in this study is International Corpus of English (ICE) (Greenbaum & Nelson, 1996; Kirk & Nelson, 2018) which comprises of several corpora from different countries. This data-set was built with the intention of providing a resource to conduct comparative studies of English used in different countries where English is the native language or the second language. In order to maintain consistency among each country corpus, common corpus design, corpus size and a common scheme for grammatical annotation have been followed (Kirk & Nelson, 2018). This data-set consists of both written and spoken English texts. Only written texts have been considered for this study. Written texts were gathered from many areas, such as; student writings, letters, academic writing, news reports, instructional writing, persuasive writing and creative writing (Kirk & Nelson, 2018). The authors and speakers of the texts are aged 18 or above, educated through the medium of English, and were either born in the country in whose corpus they are included or spent the majority of their lives there, or moved there at an early age, and received their education through the medium of English in the country concerned.

Written English corpora from Sri Lanka, India, Philippines, Singapore, Canada, Hong Kong, Nigeria, Ireland, Jamaica and USA have been collected and used for this research. Each country is consisted of 200 text documents (2,000 words per document). As Figure 1 (Right) depicts after pre-processing and normalization steps; average text document count per country is increased to 1,178 (on average 378 words per document).

3.2 DATA ANALYSIS

As the initial step, 30% (testing data) of the total data of each country corpus was kept as unseen or un-touched data. The remaining 70% of data (training data) was manually and further analysed using TF-IDF in order to identify the pre-processing requisites.

3.3 DATA PRE-PROCESSING

For this purpose 'strip_accents' parameter for the TF-IDF vectorizer and 'Unicode data normalize NFKD' were used. As a normalization step, documents are split into multiple documents with min-max word count per document 100-400 to avoid too short and too lengthy documents (refer Figure 1) Moreover, URLs, XML tags, e-mail addresses, censored data, line feeds were removed. The data set contained HTML entity encodings and those were decoded.

3.4 CLASSIFICATION MODEL BUILDING

The training data set were used for model building, parameter tuning, training and selection. This labelled training data was tokenized using the TF-IDF vectorizer. Vectorized output was fed to ML classifier to train the model. Further, k-fold (k=5) cross validation technique was used on the training data. Testing data set was kept hold to feed and evaluate the classification model in later stages. Models were built based on two classification strategies;

- **Flat Classification Strategy**

Flat classification refers to a single classifier at the root level as the decision point as depicted in Figure 2(Top). This classifier will handle the all classes as per its classification approaches such as one-vs-rest, one-vs-one...etc. In this study 'one-vs-all' approach has been followed. Hence, the dataset was trained for 10 classes (countries) together.

- **Sub-Category Classification Strategy**

Sub-category classification employs N number of classifiers as depicted in Figure 2(Bottom Left) and Figure 2(Bottom Right). For each sub-category specialized in solving a subset of the problem in which each classifier is trained. Each sub-category was trained separately. As depicted, some are binary (marked with black boxes in Figure 2) and some are multi-class classification models (marked with grey boxes in Figure 2). All the experiments of the sub-categories which were tested are depicted in Figure 3.

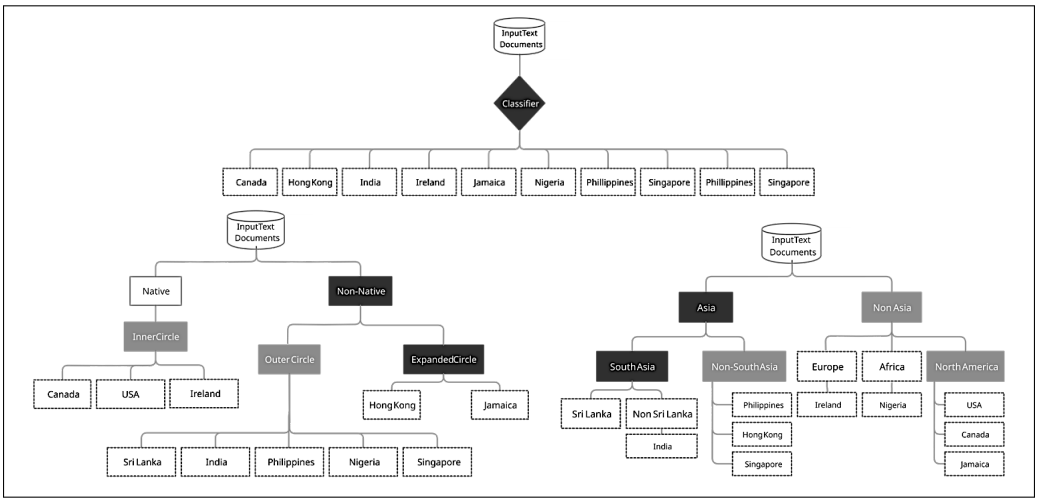


Figure 2: (Top) Flat classification strategy setup. (Bottom Left) Sub-category classification setup based on the Kachru's concentric circles. (Bottom Right) Sub-category classification setup based on the geographical distribution

	DPs/Sub-Category	Type	Description
Based on the Kachru's Theory			
1	**Native vs Non-Native**	Binary	Native (Inner circle) vs Non-Native (Outer circle + Expanded circle)
2	**Inner Countries**	Multi-class	Inner circle countries (Canada vs USA vs Ireland)
3	**Outer vs Expanded**	Binary	Outer circle vs Expanded circle
4	**Outer Countries**	Multi-class	Outer circle countries (Sri Lanka vs India vs Philippines vs Nigeria vs Singapore)
5	**Expanded Countries**	Binary	Expanded circle countries (Hong Kong vs Jamaica)
Based on the Geographical Distribution			
6	**Asia vs Non-Asia**	Binary	Asia class vs Non-Asia class
7	**South Asia vs Non-South Asia**	Binary	South Asia class vs Non-South Asia class
8	**South Asian Countries**	Binary	South Asian countries (Sri Lanka vs India)
9	**Non-South Asian Countries**	Multi-class	Non-South Asian countries (Philippines vs Hong Kong vs Singapore)
10	**Non-Asia**	Multi-class	Non-Asia class (Europe vs Africa vs North America)
11	**North American Countries**	Multi-class	North-American countries (USA vs Canada vs Jamaica)

Figure 3: Identified sub-categories based on Kachru's Theory and Geographical distribution

4 Testing and Evaluation

A combination of both 'hold-out validation' and k-fold cross-validation was used in order to reduce biasness for training and testing data set (Slotte, 2018). Hence, on the 70% of the training data set, 5-fold cross validation technique was used for model building, training and selection and 30% of the data was kept for hold-out validation to validate the machine learning model. For each decision point (where a single classifier needs to be selected) 44 experiments (4 ML x 11 n-grams) have been carried out. For each ML algorithm, word level (1-1,2,3,4) and char level (1-6,7,8,9,10,11,12) n-grams were tested. Further, performance accuracy, F1-score, precision and recall were calculated to select the best performing ML model.

5 Results and Discussion

Q1: How can texts produced by English writers in a given ethno-nationality be captured from existing corpora?

- **Flat classification strategy**

The best ML model obtained for the flat classification strategy is linear SVM trained with character 1-9 n-gram model. Furthermore, overall accuracy of 0.814 and macro average F1-Score of 0.81 was obtained. Breakdown for each country is depicted in Figure 4(Left).

In addition, Hong Kong has the highest F1-score of 1.0 while USA holds lowest F1-score of 0.73. These statistics are further verified by the confusion matrix depicted in Figure 4(Right). Accordingly, all the documents of the Hong Kong have been classified correctly. The confusion matrix depicts that model makes most of the mistakes at the classification of USA vs Canada pair. Certainly, one reason could be the geo-graphical contact between these two countries.

CLF: SVM
Analyser: Char
n-gram: (1|9)

Accuracy: 0.8140
Training Time: 193.03s
Testing Time: 3.33s

	Precision	Recall	F1-score	Support
Canada	0.8	0.78	0.79	321
Hong Kong	0.99	1	1	321
India	0.85	0.86	0.86	321
Ireland	0.78	0.81	0.8	321
Jamaica	0.83	0.79	0.81	321
Nigeria	0.78	0.79	0.79	321
Philippines	0.81	0.82	0.81	321
Singapore	0.85	0.71	0.77	321
Sri Lanka	0.81	0.79	0.8	321
USA	0.67	0.79	0.73	321
Macro avg	0.82	0.81	0.81	3210
Weighted avg	0.82	0.81	0.81	3210

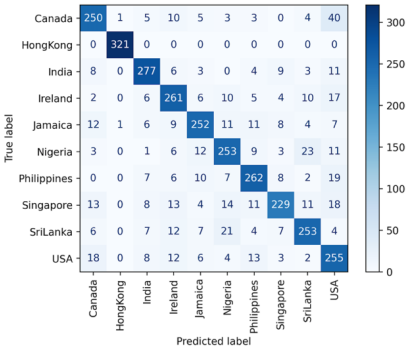


Figure 4: (Left) Test Results for the Flat Classification Strategy. (Right) Confusion matrix for the Flat Classification Strategy

These countries also belong to same inner circle according to Kachru's theory as well. Interestingly, there are high number of mistakes for Nigeria vs. Sri Lanka pair as well.

- **Sub-category classification strategy**

In sub-category classification, each sub-category can be perceived as a Decision Point (a single classifier) in the hierarchy of a decision tree. Figure 5 (Left) depicts the test results of the selected ML models for each sub-category on the training data. Accuracies of different sub-categories have varied in between 0.7549 1.000 and F1-score on the same, remains between 0.75 1.00. Based on the Kachru's concentric theory, lowest F1-Score of 0.75 is for "Native vs Non-Native" sub-category. On the other hand, 'Expanded' sub-category (Hong Kong vs Jamaica) has the highest F1-score of 1.00. In geo-graphical distribution-based model, lowest F1-score of 0.84 is for "Asia vs Non-Asia" sub-category, while 'South Asian countries' sub-category holding highest F1-score of 0.96.

	Decision Point	CLF	Analyzer	N-gram Range	Avg Train Time	Avg Test Time	Accuracy	Precision	Recall	F1-score
Based on the Kachru's Theory										
1	Native vs Non-Native	MNB	Word	(1| 2)	0.05s	0.02s	0.7549	0.77	0.75	0.75
2	Inner Countries	SVM	Char	(1| 6)	3.95s	0.10s	0.8653	0.86	0.87	0.87
3	Outer vs Expanded	SVM	Char	(1| 6)	5.25s	0.04s	0.9023	0.91	0.9	0.9
4	Outer Countries	SVM	Char	(1| 7)	15.59 s	0.35s	0.8386	0.84	0.84	0.84
5	Expanded Countries	SVM	Char	(1| 9)	4.85s	0.04s	1	1	1	1
Based on the Geographical Distribution										
6	Asia vs Non-Asia	MNB	Word	(1| 3)	0.24s	0.05s	0.8436	0.85	0.84	0.84
7	South Asia vs Non-South Asia	SVM	Char	(1| 8)	11.19 s	0.07s	0.9097	0.91	0.91	0.91
8	South Asian Countries	SVM	Char	(1| 6)	1.96s	0.02s	0.9556	0.96	0.96	0.96
9	Non-South Asian Countries	SVM	Char	(1| 6)	3.14s	0.08s	0.9460	0.95	0.95	0.95
10	Non-Asia	SVM	Char	(1| 7)	5.64s	0.13s	0.8621	0.86	0.86	0.86
11	North American Countries	SVM	Char	(1| 6)	3.75s	0.07s	0.8596	0.86	0.86	0.86

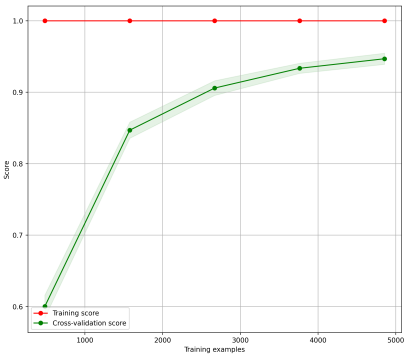


Figure 5: (Left) Test results of the sub-category classification strategy based on selected ML model. (Right) Learning curve for the Flat Classification to assess the generalisability

According to the test results most of the ML models are based on the character level n-grams. This could be due to the capabilities of character n-grams over word n-grams. Kulmizev et al. (2017) stated when compared to word n-grams, character n-grams are capable of identifying the morphological structure of a word, while the word n-grams only provides the neighbouring words. Spelling mistakes can be seen frequently in data sets; therefore, character n-grams are powerful when identifying patterns in such misspellings.

Table 2: Most significant features of the flat classification

Country	Positives	Negatives
Canada	‘canad‘, ‘”‘, ‘¡, ‘(‘, ‘)‘, ‘and‘, ‘wh‘ ‘persa‘, ‘on‘, ‘ing‘	‘’‘, ‘’s‘, ‘jamaica‘, ‘india‘, ‘ireland‘, ‘singapor‘, ‘nigeria‘, ‘philippin‘, ‘hong kong‘, ‘colombo‘
Hong Kong	‘the‘, ‘.‘, ‘ly‘, ‘s,‘, ‘t,‘, ‘hong kong‘,	‘(‘, ‘)‘, ‘”‘, ‘india‘, ‘nigeria‘, ‘philippin‘, ‘,‘jamaica‘, ‘ireland‘, ‘singapor, ‘pound sig‘, ‘sri lanka‘, ‘tamil‘
India	‘india‘, ‘on right ‘, ‘ ko‘, ‘¿, ‘the‘, ‘sing‘, ‘“‘	‘’s‘, ‘jamaica‘, ‘nigeria‘, ‘sri lanka‘, ‘singapore‘, ‘canad‘, “,‘ampersand ‘, ‘ireland‘, ‘colombo‘, ‘tamil‘
Ireland	‘’‘, ‘sign ‘, ‘ireland‘, ‘pund sig‘, ‘ampersand‘, ‘pino‘	‘jamaica‘, ‘singapore‘, ‘canad‘, ‘nigeria‘, ‘sri lanka‘, ‘hong kong‘
Jamaica	‘jamaica‘, ‘jamaican‘, ‘”‘,	‘’s‘, ‘ko‘, ‘singapore‘, ‘gap‘, ‘nigeria‘, ‘sri lanka‘, ‘canad‘, ‘colombo‘, ‘y,‘
Nigeria	‘niger‘, ‘nigeria‘, ‘n niger‘, ‘t‘, ‘ko‘, ‘r‘	‘’s‘, ‘india‘, ‘jamaica‘, ‘sri lanka‘, ‘i lan‘, ‘singapore‘, ‘colombo‘, ‘canad‘
Philippines	‘’s‘, ‘philippin‘, ‘filipino ‘, ‘the phil‘, ‘ong‘, ‘nga‘, ‘apo‘	-
Singapore	‘singa‘, ‘singapore‘,	‘nigeria‘, ‘jamaica‘, ‘singapore‘, ‘irish‘, ‘canad‘, ‘e, ‘, ‘ireland‘, ‘philippin‘, ‘india‘, ‘filipino‘, ‘pound sig‘
Sri Lanka	‘lanka‘, ‘sri‘, ‘sri lanka‘, ‘colombo‘, ‘tamil‘, ‘ko‘	-
USA	‘’s‘, ‘’‘, ‘irish‘, ‘¿, ‘(‘, ‘”‘, ‘ireland‘	‘canad‘, ‘sign ‘, ‘jamaica‘, ‘philippin ‘, ‘ampersand‘, ‘singapore‘, ‘fillipino‘, ‘pound s‘, ‘pore‘, ‘sri lanka‘, ‘ireland‘, ‘colombo‘, ‘hong kong‘, ‘right side‘

• Generalisability of the model

Assessing the generalizability of the ML models is crucial. Overfitting leads to poor generalizability of the model. Hence, in order to detect and prevent overfitting, ‘hold-out validation’ and k-fold cross-validation used. Further, learning curves on the training data set have been drawn to diagnose whether the model is overfitting or under-fitting. The red line represents the training score (trained and tested using same data) and the green line represents the cross-validation score(k=5) obtained for varying number of samples. There is a minor gap between the curves (0.95-to-1) of the learning curve as depicted in Figure 5(Right). Thus, this model does not over-fitting or under-fitting for the dataset.

• Most significant features of the model

Most significant features identified for the flat classification is depicted in Table 2. It is noticeable that when there is multi-class(N=10) classification; there is a significant biasness for each countries name and cities. Nevertheless, negative features also play a wide role in differentiating each class. However, some countries are better recognized with their currencies (i.e., Pound) and ethnic groups (i.e., Tamil, Irish).

Q2: How do texts produced by Sri Lankan English writers differ from other English writers?

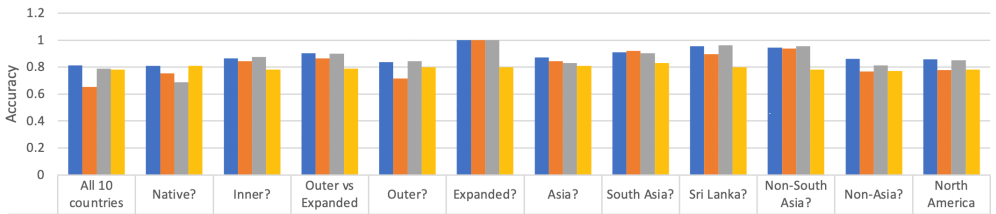
In order to analyse the significant features of the Sri Lankan (SL) English compared to native countries (USA, Canada, Ireland); significant features and grammatical phrases were derived. ’Cultural/National’ features are; Ayurveda (traditional medcine type in SL), Buddhist, sin, temple, Sinhalese, Muslims, Tamil, yala (a national park), amma (mother in English), traditional food items such as: rice, coconut, gotukola, dose and other terms such as: farmer, paddy, caste, rubber, carbide.

Main cities such as ‘Colombo’, ‘Galle’, ‘Jaffna’, ‘Kandy’, ‘Hatton’ and popular person names such as ‘Mahinda’, ‘Prabha’, ‘Rahula’, ‘Rajiv’, ‘Perera’ and popular organizations ‘LG’, ‘CEB’, ‘Seylan bank’, ‘Peoples’ bank’ also identified under Named entities. Social factors such as blasts, peace, the

LTTE, the tigers, freedom, the war, ceasefire (related to Sri Lankan war time 1983-2009), the UNP (a political party), tsunami, coal power, against women, suicide, poverty also captured.

Q3: Which machine learning techniques can gainfully employ the extracted data to identify country-of-origin of English writers?

Comparison of ML algorithms based on the selected best parameters are analysed on top of the test data and results are depicted in Figure 6 Linear SVM with SGD optimizer seems to be outperforms in most cases. Moreover, this behavior is also verified in the previous work by Kulmizev et al. (2017), Koppel et al. (2005), Ekaterina (2011), Gebre et al. (2013), Bykh & Meurers (2012). Furthermore, it is noticeable that the DT is under-performing for most cases.



	All 10 countries	Native?	Inner?	Outer vs Expanded	Outer?	Expanded?	Asia?	South Asia?	Sri Lanka?	Non-South Asia?	Non-Asia?	North America
SVM	0.814	0.8098	0.8653	0.9023	0.8386	1	0.8704	0.9097	0.9556	0.946	0.8621	0.8596
MNB	0.653	0.7549	0.8433	0.8659	0.7146	1	0.8436	0.9222	0.8972	0.9387	0.7662	0.7768
RF	0.7872	0.6865	0.8762	0.8985	0.8455	1	0.8302	0.9049	0.9625	0.9564	0.8142	0.8505
DT	0.78	0.81	0.78	0.79	0.8	0.8	0.81	0.83	0.8	0.78	0.77	0.78

Figure 6: Performance comparison of each ML technique

6 Conclusion and Future Work

With various accuracy levels the literature has proven that the ethno-nationality of a person can be identified using their written English texts and this area of research has lot of practical applications and usage. However, as discussed in the literature review still those researches are comprised with lot of limitations as stated above. Therefore, those identified limitations have reframed this research to obtain solutions to the identified research problem and defined research questions

Our work on Ethno-nationality Identification confirms that linear SVM trained with long character n-grams (1-6) to (1-9) can yield state of the art performance, and often outperform word n-grams. When determining country-of-origin it is essential to identify set of stylistic features (writing styles) which are unique to each author or a particular group of authors.

One spectacular restraint would be the usage of spelling and grammar checkers when writing English. This limitation can be overwritten when using spoken English text. Hence, as a future work spoken English text can be considered to identify the country-of-origin of the writers. Further, needs to focus more on features which have more pedagogical value. Cross corpus generalizability should be done to examine the extendibility of the model.

Ethical Considerations

The Contributors of the respective ICE Corpus countries involved, have been informed about the data collection procedures and objectives. All the identifiable named entities of the data have been anonymized in order to disjoint any relation of individuals and organizations to the data. Hence, no one, including the researchers, will be able to link data to a specific individual. Outcomes of these type of researches on categorizations of individuals base on ethno-nationality can raise concerns over racism. However, authors of this paper foresee the benefits on such categorizations over the cost.

REFERENCES

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. Toefl11: A corpus of non-native english. Ets research report series, ETS Research Report Series, 2013.

Kachru Braj B. *Asian Englishes : beyond the canon*. Hong Kong : Hong Kong University Press, 2004.

Serhiy Bykh and Detmar Meurers. Native language identification using recurring n-grams – investigating abstraction and domain dependence. In *Proceedings of COLING 2012*, pp. 425–440, Mumbai, 2012.

Andrea Cimino and Felice Dell'Orletta. Stacked sentence-document classifier approach for improving native. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 430–437, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

Kochmar Ekaterina. Identification of a writer's native language by error analysis (mphil's thesis). Master's thesis, University of Cambridge, Cambridge, United Kingdom, 2011.

Brown Eric. Native and non-native english speaking esl/efl teachers in sweden: A study on students' attitudes and perceptions towards the teaching behavior of native and nonnative english speaking teachers. Technical report, English C, 2013.

Dominique Estival, Tanja Gaustad, Son Pham, Will Radford, and Ben Hutchinson. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pp. 263–272, Melbourne, Australia, 2007.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. Improving native language identification with TF-IDF weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 216–223, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Cyril Goutte and Serge Léger. Exploring optimal voting in native language identification. In *7 Association for Computational Linguistics*, pp. 367–373, Copenhagen, Denmark, 2017. doi: 10.18653/v1/W17-5041.

Sylviane Granger. ICLE international corpus of learner english, 2014. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Sidney Greenbaum and Gerald Nelson. The international corpus of english (ice) project. *World Englishes*, 15(1):3–15, 1996.

R. Jain, Venkatesh Duppada, and S. Hiray. Seernet@inli-fire-2017: Hierarchical ensemble for indian native language identification. In *FIRE*, pp. 224–231, 2017.

Muller Jerry. Us and them: The enduring power of ethnic nationalism. Technical report, The Catholic University of America, 2008.

John Kirk and Gerald Nelson. The international corpus of english project: A progress report. *World Englishes*, 37(4):697–716, 2018. doi: https://doi.org/10.1111/weng.12350.

Moshe Koppel, Kfir Zigdon, and Schler Jonathan. Determining an author's native language by mining a text for errors. In *KDD'05*, Chicago, IL, USA, 2005.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan Noord, van, Barbara Plank, and Martijn Wieling. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 382–389, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446, sep 2018. doi: 10.1162/coli_a_00323.

Modiano Marko. *Language Learning in the Multicultural Classroom: English in a European and Global Perspective*. Studentlitteratur AB; UK ed. edition, 2009.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. Punctuation as native language interference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3456–3466, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

Jan Sapp. Race finished. *American Scientist*, 100:164, 01 2012. doi: 10.1511/2012.95.164.

Hans Olav Slotte. A shallow neural network architecture for native language identification (master's thesis). Master's thesis, Norwegian University of Science and Technology, Department of Computer Science, Trondheim, Norway, 2018.

J. Tetreault, D. Blanchard, and Aoife Cahill. A report on the first native language identification shared task. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 48–57, 01 2013.

Parham Tofıghı, Cemal Köse, and Leila Rouka. Author's native language identification from web-based texts. *International Journal of Computer and Communication Engineering*, pp. 47–50, 01 2012. doi: 10.7763/IJCCE.2012.V1.14.

ACKNOWLEDGMENT

We are deeply grateful for the International English of Corpus (ICE) project owners and maintainers of the respective countries for making available this corpus to the research community without any license fee.