DETERMINING THE ETHNO-NATIONALITY OF WRITERS 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 0.7636 for the flat classification (for all ten countries) and accuracy of 0.6224$\sim$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 word (1,1) uni-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, EFCAMDAT, 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 (EFCAMDAT), 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

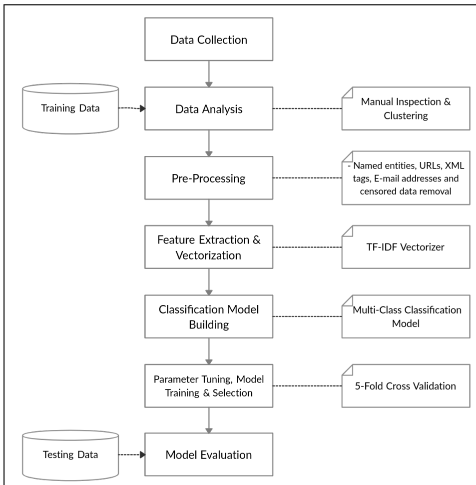
Supervised learning approach was chosen as it is suitable in classifying text documents into classes more accurately if the classes are known and the data set is labelled (Slotte, 2018). Hence, according to the literature most promising ML algorithms for the text classification such as; Support Vector 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 pre-processing, 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?

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



	Language	Total		Training (Seen Data) 70%		Testing (Unseen Data) 30%	
		#Docs	#Words	#Docs	#Words	#Docs	#Words
1	Nigeria	177	342,144	123	237,053	54	105,091
2	Philippines	178	401,871	124	281,000	54	120,871
3	Jamaica	200	412,138	140	286,651	60	125,487
4	Singapore	200	438,850	140	308,637	60	130,213
5	USA	200	428,943	140	299,509	60	129,434
6	Canada	200	427,054	140	300,106	60	126,948
7	Ireland	200	420,515	140	294,497	60	126,018
8	India	200	421,071	140	297,259	60	123,812
9	Sri Lanka	200	467,711	140	320,142	60	147,569
10	Hong Kong	200	537,153	140	375,704	60	161,449
Total		**1,955**	**4,297,450**	**1,367**	**3,000,558**	**588**	**1,296,892**

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) (refer Figure 1).

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 analyzed in order to identify the pre-processing requisites.

3.3 Data Pre-processing

Raw text-files with specific markups are used as input. These files are cleaned as per the ICE-Corpus markup guide Nelson (2002). In addition, 'strip_accents' parameter for the TF-IDF vectorizer and 'Unicode data normalize NFKD' were used. Moreover, URLs, XML tags, e-mail addresses, censored data, line feeds were removed. The data set contained HTML entity encodings and those were decoded(refer Figure 1). In order to avoid country names, nationality, currency and popular cities being trained as features, all the country specific nationality, cities and country names are removed as a pre-processing step.

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=3) 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.

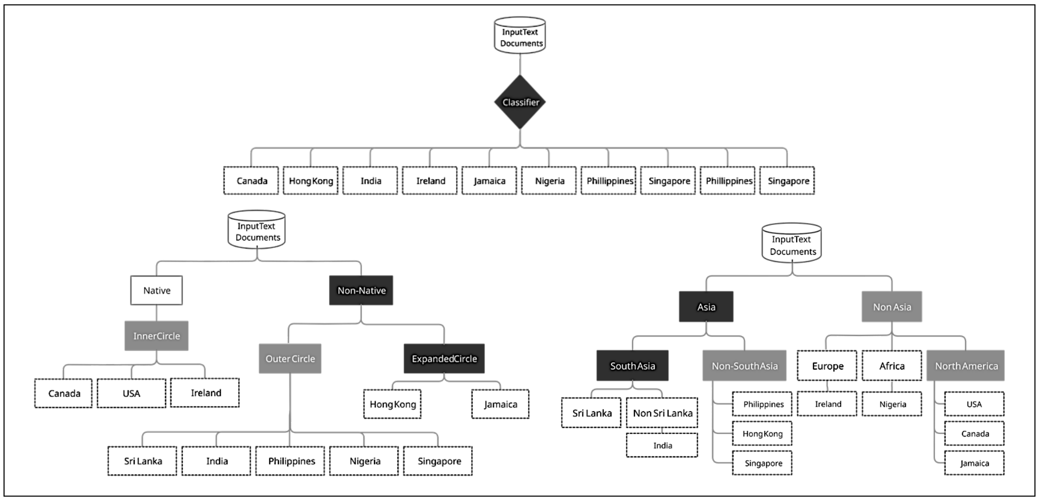


Figure 2: (Top) Flat classification strategy setup. (Bottom Left) Sub-category classification setup based on the Kachru's concentric circles. (Bottom Right) Sub-category classification setup based on the geographical distribution

	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, 3-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) 16 experiments (4 ML x 4 n-grams) have been carried out. For each ML algorithm, word level (1-1,2,3,4) 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 with SGD optimizer trained with word (1,1) uni-gram model. Furthermore, overall balanced accuracy of 0.7620 and macro average F1-Score of 0.76 was obtained. Score breakdown for each country is depicted in

Figure 4(Left). In addition, India has the highest F1-score of 0.94 while USA & Sri Lanka holding the lowest F1-score of 0.68. These statistics are further verified by the confusion matrix depicted in Figure 4(Right). Accordingly, all the documents of the India have been classified correctly. The confusion matrix further depicts that the model makes most of the mistakes at the classification of USA vs Jamaica pair. Certainly, one reason could be the geo-graphical proximity between these two countries.

CLF: SVM				**Accuracy: 0.7636**
				Balanced Accuracy: 0.7620
Analyser: Word				Training Time: 0.14s
n-gram: (1|1)				Testing Time: 0.01s
	Precision	**Recall**	**F1-score**	**Support**
Canada	0.75	0.82	0.78	60
Hong Kong	0.77	0.83	0.80	60
India	0.9	1	0.94	60
Ireland	0.73	0.82	0.77	60
Jamaica	0.73	0.75	0.74	60
Nigeria	0.83	0.7	0.76	54
Philippines	0.78	0.67	0.72	54
Singapore	0.75	0.72	0.74	60
Sri Lanka	0.69	0.67	0.68	60
USA	0.71	0.65	0.68	60
Macro avg	0.76	0.76	0.76	588
Weighted avg	0.76	0.76	0.76	588

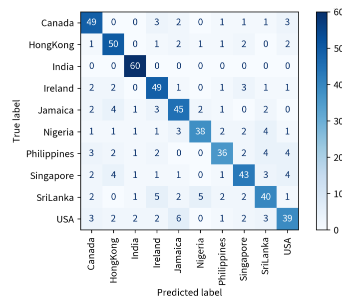


Figure 4: (Left) Test Results for the Flat Classification Strategy. (Right) Confusion matrix for the Flat Classification Strategy

• 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. Accuracies of different sub-categories have varied in between 0.6224∼1.000 and F1-score ranges between 0.49∼1.00. Based on the Kachru's concentric theory, lowest F1-Score of 0.71 is for "Outer vs Expanded" sub-category. On the other hand, 'Expanded' sub-category (Hong Kong vs Jamaica) has the highest F1-score of 0.92. In geo-graphical distribution-based model, lowest F1-score of 0.49 is for "Non-Asia" sub-category, while 'South Asian countries' sub-category holding the highest F1-score of 1.0.

	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	SVM	Word	(1| 1)	0.07s	0.00s	0.8912	0.89	0.89	0.89
2	Inner Countries	DST	Word	(1| 2)	0.73s	0.00s	0.8500	0.85	0.85	0.85
3	Outer vs Expanded	MNB	Word	(1| 2)	0.03s	0.01s	0.7672	0.81	0.77	0.71
4	Outer Countries	SVM	Word	(1| 1)	0.04s	0.00s	0.7882	0.80	0.79	0.79
5	Expanded Countries	MNB	Word	(1| 3)	0.02s	0.01s	0.9250	0.93	0.93	0.92
Based on the Geographical Distribution										
6	Asia vs Non-Asia	SVM	Word	(1| 1)	0.07s	0.00s	0.8418	0.84	0.84	0.84
7	South Asia vs Non-South Asia	SVM	Word	(1| 1)	0.04s	0.00s	0.8537	0.86	0.85	0.85
8	South Asian Countries	DST	Word	(1| 1)	0.01s	0.00s	1.0000	1.00	1.00	1.00
9	Non-South Asian Countries	MNB	Word	(1| 3)	0.07s	0.01s	0.7874	0.83	0.79	0.78
10	Non-Asia	MNB	Word	(1| 3)	0.10s	0.02s	0.6224	0.77	0.62	0.49
11	North American Countries	DST	Word	(1| 2)	0.77s	0.00s	0.8111	0.81	0.81	0.81

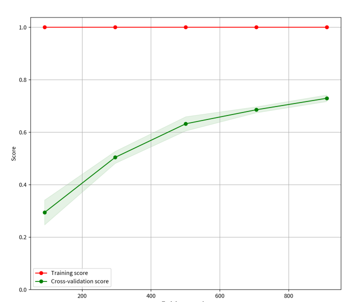


Figure 5: (Left) Test results of the sub-category classification strategy based on selected ML model. (Right) Learning curve for the Flat Classification to assess the generalisability

• Generalisability of the model

Assessing the generalizability of the ML models is crucial. Over-fitting leads to poor generalizability of the model. Hence, in order to detect and prevent over-fitting, '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 over-fitting 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=3) obtained for varying number of samples. Cross validated score curve is flattening against the training score 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 Figure 6. These features are computed base on the coefficients assigned. Some of the features are driven by the cultural and geo-graphical differences of the particular country. For instance, Ireland: northern, christmas, groundwater, queen. and Singapore: business, batik, population. Nigeria: donut(for do not), federal, god..etc.

y=Canada top features		**y=HongKong** top features		**y=India** top features		**y=Ireland** top features		**y=Jamaica** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+2.129	canadians	+2.151	government	+1.442	of	+3.445	northern	+2.756	indies
+1.966	employer	+2.101	can	+0.817	mabel	+2.388	ulster	+2.386	st
+1.944	hannah	+2.022	mr	+0.806	singh	+2.114	christmas	+2.174	mechtilde
+1.783	conventions	+1.921	body	+0.788	subhro	+2.065	cannabis	+2.113	am
+1.665	unclear	+1.682	wong	+0.732	narasimha	+1.824	sign	+1.902	clea
+1.651	prairie	+1.663	beijing	+0.722	college	+1.783	groundwater	+1.803	hi
+1.608	cegep	+1.661	japanese	+0.699	monolina	+1.668	on	+1.802	energy
+1.581	ash	+1.649	licence	+0.692	crore	+1.628	ail	+1.788	miss
+1.536	we	+1.639	mediation	+0.679	duff	+1.617	in	+1.764	seaga
+1.478	biotechnology	+1.611	yuen	+0.672	religion	+1.563	narratives	+1.760	west
+1.462	concordia	+1.574	political	+0.671	prof	+1.520	garda	+1.710	persons
+1.440	scotia	+1.528	chung	+0.670	annuity	+1.514	community	+1.632	linguistics
+1.436	cat	+1.509	ch	+0.659	books	+1.503	kay	+1.617	water
+1.411	sshrc	+1.501	kowloon	+0.658	has	+1.436	queen	+1.616	mi
+1.411	provincial	+1.468	hkac	+0.656	ayer	+1.419	bullet	+1.605	yuh
+1.405	nova	+1.446	capitalism	+0.639	earth	+1.405	jude	+1.600	crab
+1.362	ui	+1.419	diabetes	+0.632	wylie	+1.386	xmas	+1.569	campus
+1.361	information	+1.395	frank	+0.626	line	+1.346	ye	+1.568	ml
+1.349	groups	+1.394	po	+0.609	letter	+1.340	advice	+1.512	shipping
+1.347	centre	+1.378	deviant	+0.608	music	+1.324	turloughs	+1.484	mud
... 15879 more positive ...		*... 15358 more positive ...*		*... 17822 more positive ...*		*... 14395 more positive ...*		*... 14187 more positive ...*	
... 49444 more negative ...		*... 47655 more negative ...*		*... 29508 more negative ...*		*... 49248 more negative ...*		*... 49770 more negative ...*	
-1.241	road	-1.208	state	-0.499	registrant	-1.004	one	-1.141	state
-1.286	was	-1.557	this	-0.596	oil	-1.009	program	-1.156	ll
-1.405	mr	-1.574	country	-0.602	for	-1.018	<BIAS>	-1.306	it
-1.961	hope	-1.589	national	-0.944	that	-1.045	am	-1.341	ve
-6.006	can	-2.037	that	-11.279	in	-1.100	teachers	-1.343	<BIAS>

y=Nigeria top features		**y=Philippines** top features		**y=Singapore** top features		**y=SriLanka** top features		**y=USA** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+2.381	state	+1.779	labor	+1.885	business	+3.198	ltte	+2.095	that
+2.261	niger	+1.732	cattle	+1.807	batik	+1.965	plantation	+1.754	copyright
+1.891	corrosion	+1.701	la	+1.670	population	+1.956	in	+1.659	synderesis
+1.872	delta	+1.691	nog	+1.646	editors	+1.903	negotiation	+1.646	program
+1.770	kehinde	+1.654	romina	+1.584	stuart	+1.902	2000	+1.633	apartment
+1.765	facebook	+1.650	center	+1.552	curriculum	+1.896	fashion	+1.626	postmodernism
+1.750	mazi	+1.627	edsa	+1.550	tickets	+1.808	tamils	+1.508	stone
+1.739	federal	+1.610	metro	+1.525	jeremy	+1.796	unp	+1.448	kids
+1.651	donat	+1.594	differance	+1.522	hence	+1.776	ngo	+1.440	cancer
+1.606	inec	+1.508	president	+1.504	lienhwa	+1.684	ceylon	+1.432	woman
+1.605	breastfeeding	+1.490	milf	+1.497	satan	+1.658	baby	+1.355	photography
+1.585	god	+1.487	program	+1.487	client	+1.626	nanotechnology	+1.345	fp
+1.584	meaning	+1.484	inez	+1.481	phonecards	+1.617	country	+1.335	class
+1.570	handicapped	+1.484	sa	+1.476	tong	+1.598	internet	+1.333	twilight
+1.547	drying	+1.459	mindanao	+1.464	port	+1.565	tea	+1.328	sommers
+1.541	corruption	+1.439	percent	+1.458	postscript	+1.551	cholesterol	+1.323	ireland
+1.467	ekid	+1.422	ang	+1.448	national	+1.525	mudiyanse	+1.321	carbide
+1.456	governor	+1.394	levinas	+1.436	staff	+1.512	india	+1.276	europe
+1.449	urhobo	+1.393	object	+1.394	straits	+1.485	rats	+1.276	novices
+1.448	polio	+1.393	influenza	+1.391	mental	+1.449	with	+1.251	lucid
... 13779 more positive ...		*... 15933 more positive ...*		*... 15256 more positive ...*		*... 16614 more positive ...*		*... 15848 more positive ...*	
... 46418 more negative ...		*... 51205 more negative ...*		*... 48691 more negative ...*		*... 47413 more negative ...*		*... 44849 more negative ...*	
-0.928	these	-0.961	programme	-0.952	our	-0.955	must	-1.411	was
-0.970	program	-1.039	at	-1.061	people	-0.978	they	-1.690	white
-0.981	don	-1.084	real	-1.069	summer	-1.106	for	-2.080	is
-0.996	time	-1.114	labour	-1.227	says	-1.139	<BIAS>	-2.134	university
-1.057	an	-1.307	students	-1.853	we	-1.538	or	-2.899	us

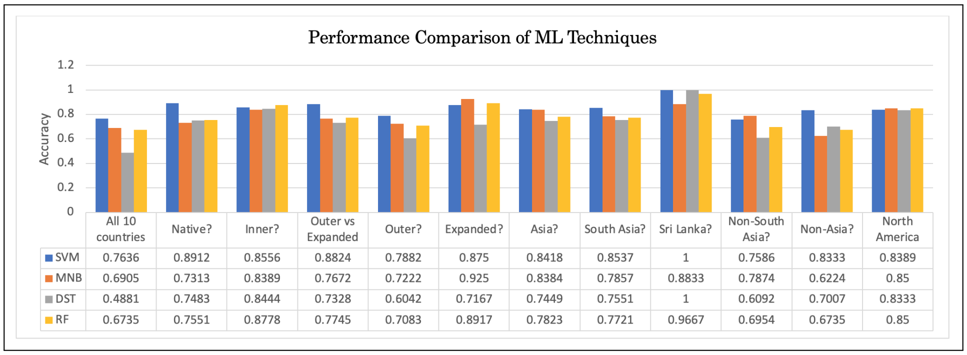
Figure 6: Most significant features of the flat classification

Q2: 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 7. 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.

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 re-framed this research to obtain solutions to the identified research problem and defined research questions.



	All 10 countries	Native?	Inner?	Outer vs Expanded	Outer?	Expanded?	Asia?	South Asia?	Sri Lanka?	Non-South Asia?	Non-Asia?	North America
SVM	0.7636	0.8912	0.8556	0.8824	0.7882	0.875	0.8418	0.8537	1	0.7586	0.8333	0.8389
MNB	0.6905	0.7313	0.8389	0.7672	0.7222	0.925	0.8384	0.7857	0.8833	0.7874	0.6224	0.85
DST	0.4881	0.7483	0.8444	0.7328	0.6042	0.7167	0.7449	0.7551	1	0.6092	0.7007	0.8333
RF	0.6735	0.7551	0.8778	0.7745	0.7083	0.8917	0.7823	0.7721	0.9667	0.6954	0.6735	0.85

Figure 7: Performance comparison of each ML technique

Our work on Ethno-nationality Identification confirms that linear SVM with SGD optimizer trained with word n-grams can yield a higher level of performance. When determining country-of-origin it is essential to identify set of features 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 transcribed text can be considered to identify the country-of-origin of the English writers. Further, needs to focus more on features which have more pedagogical value and cross corpus generalizability should be assessed to examine the extendibility of the model.

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.

Ethics Statement

The Contributors of the respective ICE Corpus countries have 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 categorization of individuals based on ethno-nationality can raise concerns over its usage and discrimination. However, author-profiling is gaining pace and the authors of this paper admire the true potential of such categorizations.

Reproducibility Statement

The implementation details this experiment is available as a supplementary material along with the submission. Kindly note that the supplementary submission does not include the complete dataset. Authors can produce the complete data-set used upon a formal request.

All the raw-data for each country in ICE corpora is available as text files. Files are pre-processed through a jupyter notebook and results were written into a csv called ‘ice-merged.csv‘. These pre-processing details are also available with the supplementary materials and for further details please refer to README.txt file in the root level.

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.

Gerald Nelson. Markup manual for written texts. Technical report, ICE, 2002.

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 Tofighı, 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.