Identifying the ethno-nationality of English bloggers using deep learning

B. S. G. Mendis

2019

Identifying the ethno-nationality of English bloggers using deep learning

A dissertation submitted for the Degree of Master of Science in Computer Science

B. S. G. Mendis
University of Colombo School of Computing
2019

UCSC

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: B. S. G. Mendis

Registration Number: 2016MCS060

Index Number: 16440602

____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms. B. S. G. Mendis

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. A. R. Weerasinghe

____

Signature: Date:

Abstract

English language is one of the most widely used languages in the world. But it is not the native language for the majority of the people. Therefore they give away certain cues such as grammar, spelling, sentence structure, frequently used phrases, etc., which separates them from native English speakers. Previous research that has been done in the area includes finding these afore mentioned cues also known as stylistic features between native and non-native English authors. These features were later used in identifying the native language of non-native English authors. However, none of those researches address the localization of English language, which occurs when a set of common stylistic features, unique to a specific locale gets absorbed by the English language. The corpora they have used are from learner corpora where authors are highly dependent on their native language. Therefore stylistic features found in written text by authors from different ethno-nationality with the same native language and cannot be identified. This research was done to uniquely identify the ethno-nationality of random authors with varied knowledge in English, based on their stylistic features, which can be used to understand the localization of English language.

The problem was addressed as a machine learning problem. Supervised learning algorithms such as Support Vector Machines (SVM), decision trees, random forest, etc. has been used before in similar work in the area. However there is a lack of using neural networks for this type of study. Neural networks have the ability to learn and model non-linear and complex relationships. It can deduce unseen relationships on unseen data which helps the model generalize and predict. This would be helpful in figuring out new stylistic features in written text. Out of many neural network models, A Long Short Term Memory (LSTM) model was chosen as it preserve the error which can be backpropagated through time and layers. There is no straight forward method to extract the stylistic features that were identified or memorized the LSTM model. Therefore as part of the research, a method involving the weights of cells in the LSTM model was developed to extract the stylistic features.

At the end of this research a model that is able to predict the ethno-nationality of an anonymous blogger with an accuracy of 62.9% was produced. In addition to that, a set of stylistic features of native authors, non-native authors, South Asian authors and Sri Lankan authors were also identified. These stylistic features can be used by linguists to further study the usage of English language.

Acknowledgement

I would like to take this opportunity to express my gratitude to everyone who helped me and guided me to make this project a success.

First of all I would like to thank the staff of University of Colombo School of Computing (UCSC) for giving me this opportunity to apply the knowledge gained through the Master of Science in Computer Science degree program.

I wish to thank my project supervisor, Dr. A. R. Weerasinghe for his helpful guidance given throughout past few months and for his excellent support.

Finally, I would like to express my immense gratitude to my family as well as all the people who helped me to successfully complete this project.

Table of Contents

List of Appendices

List of Figures

List of Tables

List of Abbreviations

CNN – Convolutional Neural Network

LDA – Latent Dirichlet Allocation

LSTM – Long Short Term Memory

NN – Neural Network

POS – Part-of-speech

RNN – Recurrent Neural Network

SVM – Support Vector Machines

Chapter 1 - Introduction

1.1 Background of the study

The oldest form of English can be traced back to 5th century and is known as Old English in the early days. It was used only by the people who lived along the North Sea coast of England [1]. However, due to British colonization in the 19th century, English language has spread across the world, making it one of the most widely used languages in the world. It is the most widely learned second language in the world and is also used as an official language in many countries [18].

Since there are many users of English language, how can one identify a person's ethno-nationality just by analyzing some text written by that person?. This can be achieved using author profiling techniques. Author profiling is a technique that is used to gain more insight about an anonymous author. It relates to both Natural Language Processing and Forensic Linguistics. It is done by analyzing stylistic- and content-based features of text.

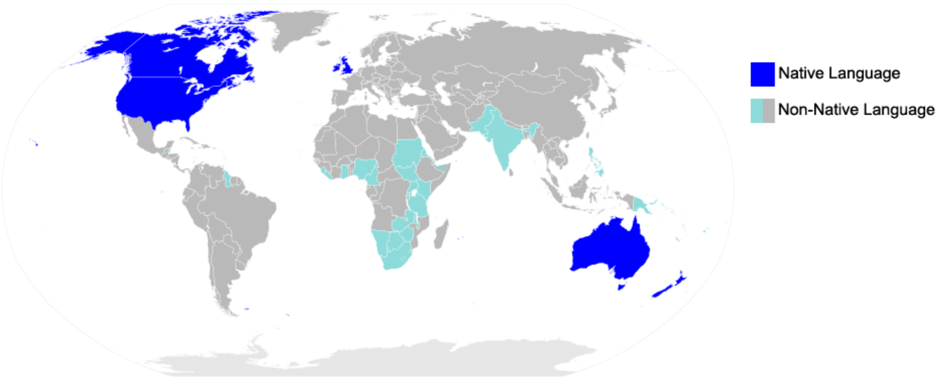

Figure 1: English Usage in the world [18]

As Figure 1 shows, for the majority of people English is not a native language. Therefore when using a language, they give away certain cues such as grammar, spelling, structure of sentences, frequently used phrases and others which separates them from native English speakers [2]. Eg:- Use of past tense to invoke politeness and indirectness (use of Could I? instead of Can I?) demonstrated by S. Biesenbach-Lucas [5] and Level of hedging demonstrated by A. Prasithrathsint [6].

Some of these stylistic features are unique to people from certain ethno-nationality. Eg:- Lack of definite article in Chinese English as explained by R. Hickey and C. Essen [12].

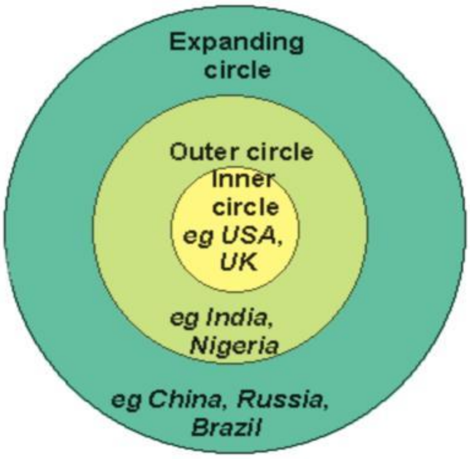

Figure 2: Kachru's 3 circles of English [13]

According to B. Kachru [13], World Englishes can be divided into 3 Concentric Circles of the language known as Inner Circle, Outer Circle and Expanding Circle as shown in Figure 2. Inner Circle includes countries where English is a native language. eg:- United States of America, United Kingdom, Canada, Australia and New Zealand. Outer circle includes countries where English was spread through imperial expansion by Great Britain and is used as a second language. In these countries English plays a historical or governmental role. Eg:- India, Sri Lanka, Malaysia, Philippines and Kenya. The Expanding circle includes countries where English is used as a foreign language mainly for international communication. English does not play either historical or governmental role in these countries. Eg:- China, Japan, Egypt and Saudi Arabia.

With the exception of Inner Circle, English language in countries belonging to other circles has gone through changes and become *localized* due to influences by linguistic, social and cultural factors of those countries.

1.2 Problem statement

Some of these stylistic features mentioned in the previous section are evident in written text as well. The question is how we can identify these stylistic features unique to each ethno-nationality and how we can use them to identify an author's ethno-nationality.

If the English language is used in an environment where there are other languages used natively by the community, features of native languages can creep into English. (Eg:- focus on subject-object-verb order). This shows that native language is not the only thing that affects the way a person uses English. If English was acquired as a second language, it affects the writer's fluency in the language. If English is taught during primary education and is also the medium of instruction in the country, it will improve the author's vocabulary which will be reflected in his writing. It also reduces the dependency the author has with his native language. If the language is used as a foreign (international) language, the localization would be less as the usage is low within the community [4].

This research will try to solve the problem of identifying the stylistic features unique to each ethno-nationality and using them to identify an author’s ethno-nationality.

1.3 Objectives of the study

The main objective of this research is to propose a new method to identifying stylistic features, unique to Sri Lankan authors who write in English, which can also be used in linguistics to understand the localization of English language. The problem will be treated as a classification problem, where countries will be considered as classes.

For this research, we will only focus on finding the stylistic features unique to Sri Lankan authors. This will be achieved by approaching the problem step by step solving couple of research questions. The research questions are:

- What feature engineering methods can be applied to an automatic author ethno-nationality identification model based on stylistic features?
- What are the stylistic features of native and non-native English authors that can be identified using such methods?
- What are the common stylistic features of South Asian English authors that can be identified using such methods?
- How do Sri Lankan authors differ in style compared to other South Asian English authors?

The hypothesis used in this research is that authors from the same ethno-nationality will share stylistic features such as choice of words, subject-verb-object order, use of prepositions etc. in their written text.

The main goals of this project includes training a model capable of identifying Sri Lankan authors based on stylistic features, developing a method to extract those features and providing a list of key characteristics found in English text by Sri Lankan authors which would be helpful for linguists.

1.4 Limitations of the study

The data required for this research will be gathered from blogs available on the Internet. Therefore false positive data that might end up in dataset used to train the model. Eg:- Text written by a British person living in Sri Lanka might be in the Sri Lankan English corpus. When creating the corpus, we rely solely on the country given by the author as his country at the time of registration. Since the data is taken via an API, there is no way to cross check the user information and verify the author's country. This can be overcome by using a large set of data where such random cases can be considered as outliers. These outliers reduces the overfitting in the model and also forces the model to learn rather than memorizing. Therefore when building the model, it was also taken into consideration.

1.5 Structure of the dissertation

This dissertation will describe the development of a neural network model that is able to predict the ethno-nationality of an anonymous blogger as well as identifying their stylistic features. Dissertation structure is as follows.

Chapter 2 - Literature Survey includes a critical analysis of previous related work done in the area of focus. Previous work on author profiling and the use of machine learning techniques are included in this chapter.

Chapter 3 - Research Methodology describes the methodology used to finalise the final model. It explains inner workings of a neural network, data collection process and the evaluation plan.

Chapter 4 - Implementation describes the process of data collection and preparation and development of the model finalized during the previous chapter. The chapter includes diagrams and tables which further explains the solution.

Chapter 5 - Evaluation and Results includes a critical analysis of the model as well as the stylistic features identified for native authors, non-native authors, South Asian authors and Sri Lankan authors.

Chapter 6 - Conclusion and Future Work describes all the closing details of the project, what lessons were learnt during the project and how the system could be further improved.

Chapter 2 - Literature Review

Many research have been done to profile authors and gain more insight about them such as automatically categorizing written texts based on gender by M. Koppel et al. [14], Predicting age by K. Santosh et al. [16]. This review will focus on, author profiling, identifying the ethno-nationality of an author and use of machine learning for author profiling.

2.1 Author Profiling

In the case of authorship attribution, researchers aim at finding a set of features that are unique for an author or group of authors. Such features include type/token ratio, average sentence and word length, part-of-speech (PoS) n-grams, syntactic rules, use of punctuation marks, conjunctions, prepositions, and auxiliary verbs [2]. Word tokens and token-based n-grams, are rarely used for this as they are more topic-specific and are useful for topic classification rather than author attribution [2].

In real-life authorship attribution problems seek idiosyncratic usage by a given author that serves as a unique fingerprint of that author [3]. Such as repeated use of particular types of neologisms or unusual foreign word usage. These stylistic features provides clues regarding the author’s age, gender, native language, country and etc. According to Koppel et al. automatic text analysis based on function words and PoS n-grams can identify an anonymous author’s gender with an accuracy of approximately 80% [14].

By using content based features, style based features and topic based features, K. Santosh et al. [16] were able to predict both the gender and age of the author with an overall accuracy of 54.8% to analyze how everyday language reflects basic social and personality traits. Ngrams, POS (part-of-speech) tags, LDA (Latent Dirichlet Allocation) Topic Model and Scores of classes from different models were analyzed using machine learning algorithms such as SVM, MaxEnt and Decision Tree.

2.2 Identifying the ethno-nationality

A closely related area of this research includes identifying the native language of non-native English speakers. For this, separately identifying native speakers from non-native speakers is essential. There are many features in the text written by non-native speakers which can be used for this purpose. Eg:- grammar, spelling, structure of sentences, frequently used phrases and etc. [2, 12]

Biesenbach-Lucas examined of e-politeness among native and non-native speakers of English by examining emails sent by students to professors [5]. Emails written for three different scenarios such as requests for appointment, requests for Feedback and requests for extension of due date were analyzed for this research. Politeness features used in this research includes syntactic modifiers and lexical modifiers. Syntactic modifiers includes use of past tense ("Could I" instead of "Can I"), progressive aspect ("I was wondering" instead of "I wonder") and embedding ("I would appreciate it if you could…"). Lexical modifiers include the usage word please, understaters, subjectivizers, consultative devices and hedges. After analyzing the emails it was found that both native speakers and non-native speakers tended to use the same general strategies but with different types of politeness devices [5]. Which points toward lack of linguistic flexibility and idiomatic expressions, unawareness of letter conventions transferable to email, and inability to select appropriate lexical modification among non-native speakers.

Mitigating words so as to lessen the impact of an utterance can also be regarded as an important feature in English writing [6]. It can be analyzed to identify native speakers and non-native speakers. Prasithrathsint [6] analyzed the use of hedging in English academic papers written by native and non-native speakers of English, using text collected from countries each belonging to one of Kachru's three circles of English. The research shows that the hedging devices are used most by native speakers (7.6 times per 1000 words of text length), the second most by Filipino scholars (6.6 per 1000 words of text length) who also belong to the Outer Circle, and least by Thai scholars (3.5 per 1000 words of text length) who belongs to the Expanding Circle. It is evident that the frequency of hedging in academic writing in the humanities depends on the degree of native competence in English [6].

The above research clearly shows the impact a language acquisition mode has on an author. The internationalization of English has prompted the diversification of English [4]. Each country uses the language in its traditional cultural and linguistic contexts. Which allows a distinct version of the language to be created with unique structural and functional features. These features can be helpful when trying to identify an author's native language, ethnicity, nationality and etc.

2.3 Use of Machine Learning for Author Profiling

M. Koppel et al. [3] used error analysis to identify an author’s native language. For this they used SVMs, which is also one of the highly used supervised learning algorithms used for text classification. SVMs can handle infinitely many features and the only thing that it needs compute efficiently is the similarity of two examples. SVM does not need an aggressive feature selection. There are formulas, which can help in predicting how good a classification of an unseen example will be by error-estimating. This eliminates the need for cross validation techniques [19]. Errors such as Orthography, Syntax, Neologisms and Parts-of-speech bigrams were analyzed using a multi-class linear Support Vector Machines (SVM). Koppel and team managed to successfully implement a fully automated method for determining the native language of an anonymous author. The experiments were performed on a corpus including authors from five different countries, Czech, France, Bulgaria, Russia and Spain. The method achieved an accuracy of above 80% in categorizing unseen documents [3]. The data used in this research was taken from the International Corpus of Learner English. Therefore the dependencies authors have with their native language is high.

E. Kochmar [2] also used error analysis to identify an author’s native language. A number of binary classification experiments with Support Vector Machines (SVM) have been carried out in this research. Apart from that, a distributional analysis which seeks to identify specific patterns in the use of English was also performed based on an assumption that linguistic properties are distributed differently in texts produced by native speakers of different languages [2]. The results of this research shows that languages of two Indo-European language groups, namely Germanic and Romance, can be distinguished with an accuracy of 84.35%. In the data, the classification accuracy for the language pairs within these branches ranges from 68.40% for the Spanish – Catalan pair to 100.00% for the Danish – Swedish pair. In this research also the corpus used is a learner corpus, which is The Cambridge Learner Corpus. Therefore the dependencies authors have with their native language is high.

In Estival et al., five demographic and five psychometric traits, namely age, gender, native language, level of education and main country of residence for the demographic traits, and agreeableness, conscientiousness, extraversion, neuroticism and openness for the psychometric traits were considered as characteristics of anonymous authors [15]. After using multiple combinations of machine learning algorithms and features, which gives us the best result for each trait, an accuracy of 84.22% was obtained for the predictions made for the trait native language while an accuracy of 81.13% was achieved for the trait country. Machine learning algorithm random forests were used in this research. Random forest is flexible and has the ability to be used as both classification and regression tasks [20]. Corpus used in this research contains a wide variety of people and not solely students. Therefore the dependency authors have with their native language is low.

According to Hickey et al. localized varieties of English language can be found in countries belonging to both Inner Circle and Outer Circle [12]. Caribbean islands, South Africa and Zimbabwe belongs to the Inner Circle. Yet they show a clear phonological and grammatical derivation from varieties in Britain and Ireland as English has gone through continuous transmission due to emigration and settlement at overseas locations. Outer Circle varieties can be found mostly in Asia and Africa and easily recognized by their phonologies. In China, many varieties of Chinese languages such as Mandarin, Xiang, Yue and etc. are spoken. In Chinese English, there exists very prominent grammatical features such as the absence of the definite article, the lack of inflectional endings and different application of prepositions. These features are clearly the result of multiple Chinese languages that exists in the background [12]. This supports the fact that an author’s native language alone does not affect his or her writing style but also the general usage of English language in the author’s surrounding environment.

Marco Lui et al. [21], classified English documents by national dialect. In this research lexical and syntactic variation between English dialects of Inner Circle Countries Australia, Britain and Canada were taken into consideration. The text sources were collected from, national corpora, open web, government web and twitter. This balances the demographic aspects of the authors throughout the corpora. After analyzing the text using an SVM, it was evident that the English dialects have systematic differences at the syntactic level. These characteristics go beyond simple topical differences, as representations such as function word distributions, and part-of-speech plus function word bigrams, omit topical information from consideration [21].

Author profiling for English emails was done by Estival et al. which focused on providing probabilities for the author’s basic demographic traits gender, age, geographic origin, level of education and native language. Data used in this project was the same they used in their Text Attribution Tool (TAT) project. Country of origin in this case had four classes, United States of America, Egypt, Australia/New Zealand and Other. The machine learning algorithms used include decision trees, random forest, lazy learners, rule-based learners, SVMs, LibSVM, ensemble/meta-learners and AdaBoostM1. These algorithms were used in combination with feature selection methods [22]. The results of this research shows that combining multiple machine learning algorithms is successful for binary as well as n-ary classifications on very diverse classification tasks.

2.4 Deep Learning for Text Analysis

Machine learning has two types of learning techniques known as supervised learning and unsupervised learning. Supervised learning is used to build models that make predictions based on classification and regression. Unsupervised learning is used to find hidden patterns or intrinsic structures in data. Since this research requires a predictive model that can be used to predict an author's ethno-nationality, supervised learning will be the appropriate approach.

There are number supervised learning algorithms such as Support Vector machines (SVM), Naive Bayes, Decision Trees, Neural Networks etc. Eg:- M. Koppel et al., 2005 used SVMs [3] and Estival et al., 2007 used decision trees [22]. Out of these methods, neural networks has the ability to learn and model non-linear and complex relationships as in real life. Also after learning from the initial inputs and their relationships, it can deduce unseen relationships on unseen data which helps the model generalize and predict [27].

Neural Networks or Artificial Neural Networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains [24]. As shown Figure 3, Neural Networks consist of input and output layers, as well as one or more hidden layers which extracts different levels of information and transform the input. They are designed for detecting patterns in data. They can be used for tasks such as classification, clustering and prediction. Similar to humans, neural networks learn from experience. For neural networks that experience comes from the training data. The more data it has, the more accurate it will become.

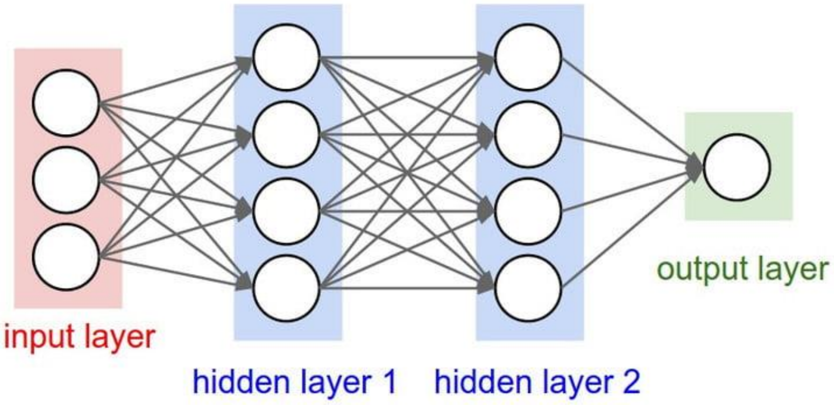

Figure 3: Artificial Neural Network [25]

2.5 Conclusion

After reviewing the literature in the previous section, it is evident that though many machine learning algorithms have been used for research, there is a lack of using deep learning to solve similar matters.

The localization of a language and the effect of its surrounding environment has on the language itself can be studied using a demographically well balanced data set. Text by authors with various levels of competency in English will be helpful to lower the dependency the authors have with their native language. Since the corpus used in this research is not a pre-tagged corpus, neural network's ability to learn is essential. This can be used to identify the stylistic features unique to each ethno-nationality.

Chapter 3 - Research Methodology

Based on the hypothesis that authors from the same ethno-nationality will share stylistic features such as choice of words, subject-verb-object order, use of prepositions etc. in their written text, the problem of identifying stylistic features unique to each ethno-nationality and using them to identify an author’s ethno-nationality will be solved in this research.

This chapter will focus on how the proposed solution will be implemented and tested as well as the methods used to collect data.

3.1 Research design

This research consists of three key activities. Data collection, Implementing and training the neural network and extracting the features.

The first step is to collect the data that is necessary to build a corpus. Once the data is ready, the neural network model can be implemented. Implementation and training of the model will happen simultaneously as the model's design and accuracy depends on each other. The model will be trained to predict the author's ethno-nationality based on stylistic features.

After training the model, it will be validated using a set of test data. Stylistic features which were identified by the model will be extracted by analysing the weights assigned for each word in cells of neural network's layers.

3.1.1 Data collection

In order to find out these common stylistic features, informal text written by both native and non-native English speakers was analyzed. Formal text such as dissertations, research papers, official letters, newspaper articles etc. was not analyzed in this case as the probability of somebody with a greater fluency of English than the original author, proof reading and making corrections is high. This would prevent common grammatical errors, frequently used terms, words etc. to disappear from the text. Learner corpora of English also was not used as such authors have a higher dependency with their native language.

Due to the reasons mentioned above the informal text, we have chosen to analyze blogs written by both native and non-native English speakers. This approach allows us to analyze text by authors coming from various backgrounds. Google Blogger provides an easy to use Javascript based API which can be used to filter blog posts based on an author's country as well as written language. Therefore the required data was gathered using Google Blogger API [54]. Personally Identifiable Information (PII) will be removed from the posts to protect author's privacy. Blog posts from countries belonging to each of the three circles of English were collected for this. Therefore blog posts from United States of America (USA) and United Kingdom (UK) were collected to represent the Inner Circle, posts from Sri Lanka, India, Pakistan, Kenya and Malaysia were chosen for the Outer Circle while China and Japan were chosen for Expanding Circle.

First a list of author IDs based on their country was derived. Then using those author ids, a list of blogs written by them in English and their blog ids were collected. Then a set of blog posts from the said blogs were retrieved using the API. The post content comes as an HTML element when retrieved using the API. Therefore the content was cleaned up by removing unnecessary HTML tags using JavaScript. Then the post content was saved into an XML file.

3.1.2 Implementing the neural network

Text Analysis is the process of parsing texts in order to extract machine-readable facts from them. Its' purpose is to create structured data out of free text content. The process can be thought of as slicing and dicing heaps of unstructured, heterogeneous documents into easy-to-manage and interpret data pieces [23].

There are many types of neural networks such as Feedforward Neural Network, Radial basis function Neural Network, Kohonen Self Organizing Neural Network, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Modular Neural Network etc. All of these networks are named after the way they pass information through a series of mathematical operations performed at each node of the network.

Out of these methods, Recurrent Neural Networks (RNNs) shown in Figure 4, behave more like a human brain by saving the output from previous state in memory and reusing it as an input in the current state.

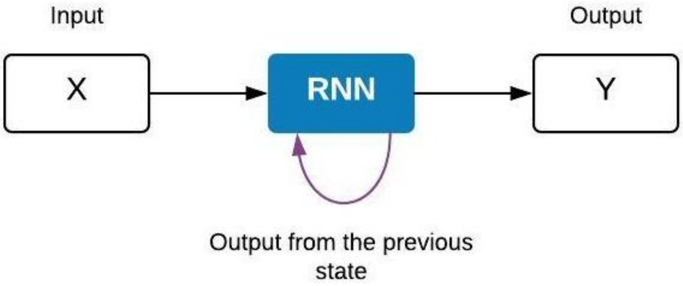

Figure 4: Recurrent Neural Network

There is information in the sequence itself. By adding memory to neural networks, recurrent nets use it to perform tasks that feedforward networks cannot [26]. Due to backpropagation, each hidden state contains traces not only of the previous hidden state, but also of all those that preceded it for as long as memory can persist [25].

As most neural networks, recurrent nets also faces the vanishing gradient problem. This reduces the learning ability of the network. As a solution for this, Long Short-Term Memory Units (LSTMs) were introduced.

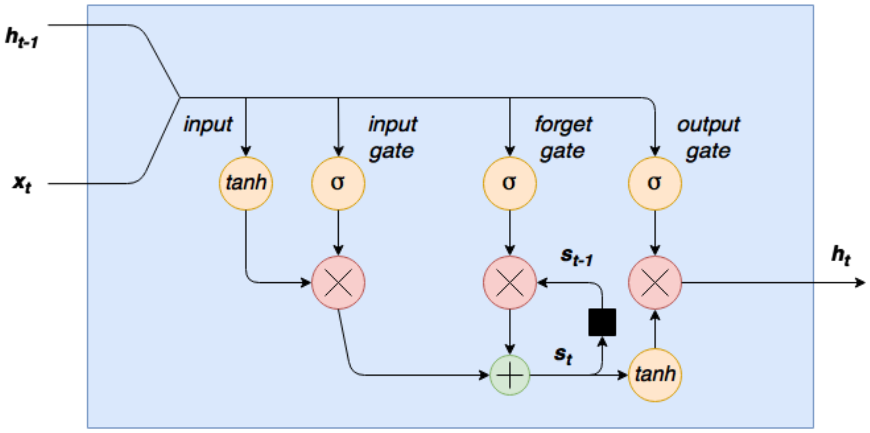

Figure 5: LSTM cell diagram [28]

Figure 5 shows a LSTM cell diagram. LSTMs preserve the error which can be backpropagated through time and layers. Recurrent nets to continue to learn over many time steps (over 1000) by maintaining a more constant error. Which thereby opens a channel to link causes and effects remotely [26]. LSTMs achieves this by using a set of gated cells. These cells while acting as memory decides what data to store and when they should be read or deleted. Due to this feature in LSTMs, what was learnt by the model does not get destroyed by the new inputs. Therefore LSTMs would be able to remember the features identified in text belonging to each class for a long period of time thus making it a suitable model for this research.

3.1.3 Extracting stylistic features

Neural networks are good at making predictions. Yet the process of making that prediction is not visible to the user. LSTM neural networks contains LSTM layers. These layers contains cells that store information as weights. In this case the cells will store a weight for each word in the provided text.

To extract the features, sample text will be run through the model and instead of the output of the final dense layer, we will be taking the output of the LSTM layers in the model as well. The output of these layers will be the weights stored in the cells for each word. From each layer, cells that contains an above average weight will be selected. Then they are mapped with the corresponding words in the sample text to find the stylistic features.

3.2 Evaluation plan

After a successful training, the model should be able to correctly distinguish Sri Lankan authors from foreign authors. The test data set will be used for this process. Since this is a classification based model, there are four types of outcomes that can occur.

True positives - when you predict an observation belongs to a class and it actually does belong to that class.
True negatives - when you predict an observation does not belong to a class and it actually does not belong to that class.
False positives - when you predict an observation belongs to a class when in reality it does not.
False negatives - when you predict an observation does not belong to a class when in fact it does [30].

Based on the above, the accuracy, precision and recall of the model can be measured as below.

Accuracy = correct predictions / all predictions
Precision = true positives / (true positives + false positives)
Recall = true positives / (true positives + false negatives)

LSTMs sometimes become underfit or overfit depending on the test data used. To make sure the model is behaving correctly, the performance of the training data set and the validation data set will be plotted. If the model is working correctly, the train and validation loss decrease and stabilize around the same point as shown in Figure 6.

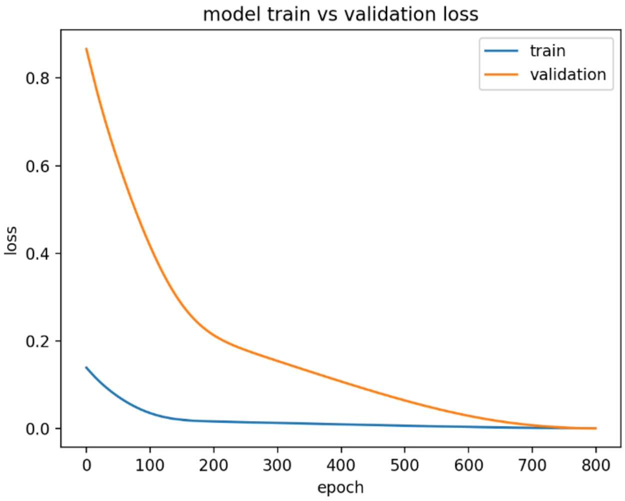

Figure 6: Diagnostic Line Plot Showing a Good Fit for a Model [29]

Chapter 4 – Implementation

This chapter focuses on the implementation of the solution explained in Chapter 3. Data Collection and preparation, implementation of the neural network and extracting stylistic features will be further discussed here.

4.1 Data Collection and Preparation

For this analysis, text extracted from Google Blogger will be used as data. The dataset contains 14189 labeled posts. Data was collected using the Google Blogger API [54]. First a list of author ids based on their country was derived. By using those author ids, a list of blogs written by them in English and their blog ids were collected. Blog posts from the said blogs were retrieved using the API by providing the blog id. The content of the post is retrieved as an HTML element when retrieved using the API. Therefore the content was cleaned up by removing unnecessary HTML tags using JavaScript. Then the post content was saved into an XML file. All the posts collected using this process has more than 100 words in each.

Still there were some inconsistencies with data. Mainly caused by transliteration. This was evident in some of the blog posts collected from Sri Lankan authors.

Eg:- A blog post taken from a blog written by a Sri Lankan author [47].

<post>
Isellaama New project gihin form ekak ganna.dan eeka kamati size ekakata drag karala hadaa ganna. (ex:- 4800*3600)Dan Text Box button eka "General Tab" eken toorala MS paint ekedi kalaa wage Form eka atule andaganna.Dan ee wagema command button ekat andaganna.OKDan api Form eka design karagena iwaray.Dan tiyenne programme karanna.baya wenna deyak naa....meeka leesi wadak !Dan api programme karanna yanne , "Hello" kiyana button eka ebuwama , programme eke text box eke " Hi , Kohomada ? " kiyala watenna.eekata api issellaama coding window eka open karaganna ooni.apita ooni Command button eka ebuwama reaction eka wenna nisa api Command button eka double click karanna ooni !Dan mee wage window ekak open wenna ati.Ooke tamaa api programme karanne.mataka tiyaaganna ooni deyak tiyenawa, ee tamaa VB6 wala coding gahala save karanna deyak naa . (coding window eka) eeka nikanma save wenawa! codings gahala iwara wela nikamma coding window eka close karala daanna.command button eka double click kalaa ma oya gollanta dakinn labey mee wage lines tikak ..Private Sub Command1_Click() End Subooke command button eke prperties window eken Name kiyana eka wenas kalaanam command1 wenuwata ee daapu name eka dakinna labey...api command button ekak hari , form ekak hari , text box ekak hari mokak double click kalat apit oya wage code line dekak dakinna puluwan...api dan gahana coding okkoma gahanna ooni oya lines deka atule. (advanced programming waladi eliyet gahanna puluwan)dan ayet codeing gana balamu..apita dan ooni command button eka click

kalaama text box eke "Hi , kohomada" kiyala watennane ... ee nisa api gahanawaPrivate Sub Command1_Click()Text1.Text = "hi , kohomada "End Subtext box eke Name eka wenas wale natnam Text1 kiyalama tiyanna.wenas kalaa nam ee nama daanna. (ex:- changedname.text)dan tikak hari terenna ati ne .... dan aye ayet mee wage podi programmes hada hadaa balanna..... wenin wenin massages daanna.Prashana ?Prashane - text1.text ! mokatada text1 kiyala gahala .text kiyalat gahuwe ?Answer eka - ee hati , api text box ekak gahuwama .text kiyalat gahanna ooni.Label ekak nam .Caption .. meewa issarahadi teerey !Download This project*To Download Right click and select Download*1.4 kb*compressed using WinRAR*Befor use right click on downloaded file and select "Extract"* Need Help ? contact Web Master
</post>

Though the English language was selected when retrieving blog ids, blogs written in other languages which uses the Latin alphabet was also included in the data set.

Eg:- A blog post taken from a blog written by an American author [48].

<post>
001154852CCVI72100414.0 Normal 021 false El 22 de abril, Día de la tierra, Caritas del Perú con apoyo de CEAS y RED MUQUI lanzaron a nivel nacional: “Seamos profetas de la vida. Protejamos nuestra tierra, nuestra casa común”. Una campaña para cuidar la creación de Dios. Y es que los Mensajes de Papa Francisco a los Movimientos Populares, nos llaman a valorar a nuestras comunidades campesinas, sus formas de vida respetuosas y defensoras de la naturaleza, de sus tierras, de sus recursos naturales, que son de los que todas/os vivimos, nos alimentamos, los que nos dan la seguridad alimentaria, y por ello es necesario proteger sus tierras, que son parte de nuestro planeta tierra, de nuestra casa común... El Papa Francisco nos llama en su Encíclica “Laudato Si” a todas/os a respetar nuestra tierra, nuestro planeta, nuestra casa común, para que tengamos un futuro seguro y sostenible donde los más pequeños/as, nuestros hijos/as puedan vivir plenamente. Por ello: “Seamos profetas de la vida. Protejamos nuestra tierra, nuestra casa común”. Puedes descargar la Campaña aquí: urlLink Primera acción día de la tierra by Katty Huanuco on Scribd
</post>

To overcome these problems mentioned above, all of the posts were filtered using langdetect 1.0.7. A language detection library adapted from Google's language-detection.

After cleaning up the data to remove unnecessary HTML tags, the collected data was divided into two sets of equal sizes. One data set was used for training the model while the other was used to test it. Remaining data processing tasks such as tokenizing was done within the model.

Machine learning techniques have been widely used for text analysation. The goal of this research is to create a model that is capable of analyzing unstructured text, learn the patterns and make predictions with respect to an author’s ethno-nationality. Therefore it was decided to build an LSTM model. LSTMs falls under Recurrent Neural Network (RNN), which is a supervised learning technique.

4.2 Implement the Neural Network

To build the model, python was used as the main coding language along with multiple python libraries. The libraries and their usage is explained in Table 1.

Library/Package	Usage
Pandas [31]	Provides easy-to-use data structures which is used when reading the training and test data CSVs.
Numpy [32]	Is the fundamental package for scientific computing with Python. Here it is used to calculate the length posts, mean of accuracy and loss etc.
Os [33]	Used to read, write, open or close files.
Tensorflow [34]	Used as the backend of the model
Nltk [35]	Natural Language Toolkit (NLTK) is used to work with human language data.
Re [36]	Used to handle regular expressions.
Time [37]	Used for various time-related functions
sklearn.model_selection -> train_test_split [38]	Used to split the training data set into training and validation data set
keras.preprocessing.text -> Tokenizer 39]	Used to vectorize the text corpus, by turning text into a sequence of integers
keras.preprocessing.sequence -> pad_sequences [40]	Used to Pads sequences to the same length.
collections -> namedtuple [41]	Used to create data structures that have fields accessible by attribute lookup as well as being indexable and iterable.
tensorflow.contrib.tensorboard.plugins -> projector [42]	Used to visualize data

Table 1: Python Libraries and their usage

The first task which is normally done during text analyzation is to remove stop words such as ‘the’, ‘is’, and ‘are’. In this model stopwords will not be removed because during the literature survey it was learnt that the usage of stopwords can also be considered as feature. Eg:- Level of hedging [6] and lack of definite article in Chinese English [12].

Once the data has been loaded, the next step is to tokenize it. Tokenizing is the process of converting words into numbers. Eg:- ["The”, “quick”, “brown”,“fox” ,“jumps” ,”over”,“the” ,“lazy” “dog”, “.”] it would be tokenized to [1, 2, 3, 4, 5, 6, 1, 8, 9]. The total vocabulary of this project is 179724. When tokenizing, each of those 179724 words receives its own, unique number. The next task is to make all of the reviews the same length. Though all the posts have more than 100 words in each, they vary in length. Longer posts with more words should improve the accuracy of the model. However to increase the training speed of this model we have limited it to 200. Using pad_sequences extra words in posts with more than 200 words will be removed while posts with less than 200 words will have padding tokens added until it reaches a length of 200.

After padding, the training data set is divided into training and validation sets. Validation set would 0.10 of the total training data set. A random state of 42 was used to shuffle the data. Therefore different train/test split will occur in each run of the model.

The next task is to implement the model. A high level structure of the model is shown in Figure 7. The purpose of each layer will be discussed in the next section.

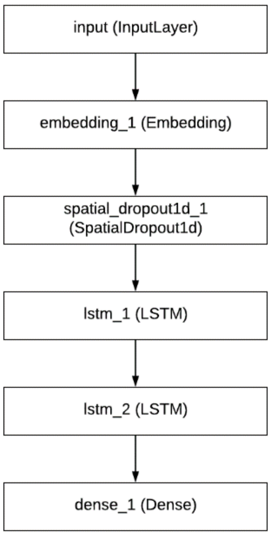

Figure 7: High Level structure of the model

The basic structure of the model is as below.

```
Layer (type)                     Output Shape            Param #
=================================================================
input (InputLayer)               (None, 200)              0
_________________________________________________________________
embedding_1 (Embedding)          (None, 200, 200)         35944800
_________________________________________________________________
spatial_dropout1d_1 (Spatial     (None, 200, 200)         0
_________________________________________________________________
lstm_1 (LSTM)                    (None, 200, 100)         120400
_________________________________________________________________
lstm_2 (LSTM)                    (None, 100)              80400
_________________________________________________________________
dense_1 (Dense)                  (None, 9)                909
=================================================================
Total params: 36,146,509
Trainable params: 36,146,509
Non-trainable params: 0
_________________________________________________________________
```

Figure 7 shows the high level structure of the model. The model consists of 6 main layers. The Input Layer, Embedding Layer, Spatial Dropout layer, 2 LSTM layers, and a Dense layer. The first layer is the input layer. The second layer is the embedding layer. It is initialized with random weights which will learn an embedding for all the words in the training dataset.

The second layer is the SpatialDropout1D layer. This layer drops entire 1D feature maps instead of individual elements. If adjacent frames within feature maps are strongly correlated, regular dropout will not be able to regularize the activations. This will otherwise just result in an effective learning rate decrease [44]. A dropout of 0.5 is used in this model.

The third, fourth and fifth are the three LSTM layers used in this model. Each of them 100 cells inside them and 0.6 dropout is also used. The final layer is a Dense output layer. The dense layer used in this model is capable of handling 9 different classes.

During training, RMSprop algorithm was used as an optimizer. RMSprop (root mean squared prop) root mean squared prop lies in the realm of adaptive learning rate methods, which have been growing in popularity [45]. It deals with radically diminishing learning rates.

For predictions, Softmax activation function used as it can handle multi class problems. The output of the softmax function is a categorical probability distribution, which tells the probability that any of the classes are true [46].

4.3 Extracting Stylistic Features

In order to analyse the stylistic features, the weights assigned by the model for each cell in each LSTM layer was analysed. The weights will be analysed when making predictions using the model. When making predictions, the model's output is limited to the final dense layer. We can see only the class that it has identified as the output. Eg:- US, UK, SL. Therefore we need to expose the outputs of the 2 LSTM layers in the model.

First we expose the output of the outermost layer of the LSTM layers, which is LSTM_2 layer as shown in Figure 8. This output shows the weights stored in each cell of the LSTM layer.

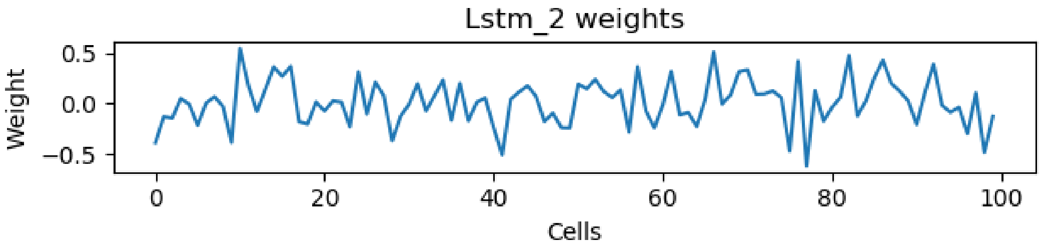

Figure 8: Weights stored in each cell in LSTM_2 layer

From this layer, we select a set of cells with an above average weight. These are the cells that stores information with significant value. Therefore their weights are higher than average. Next we expose the output of LSTM_1 layer. This layer gives the weight, each cell has stored for each word in the sequence as shown in Figure 9.

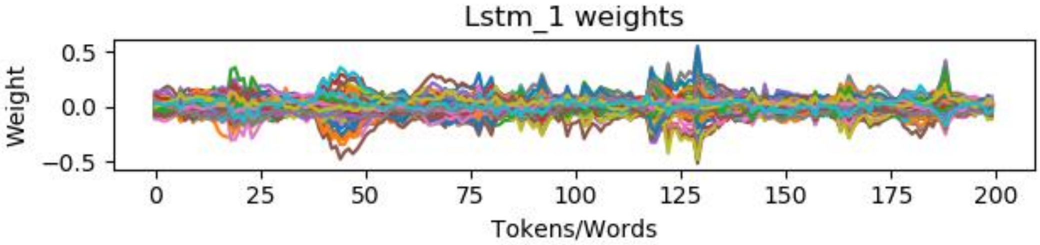

Figure 9: Weights of each word in LSTM_1 layer

Here an assumption is made that the cells which has a higher weight will continue to keep a higher weight in the following layers if the particular word has a significance. For each word, the maximum weight along with the relevant cell is selected from LSTM_1 layer. Once the maximum weight for each word has been identified, the words with an above average weight is selected. Thus filtering the words with significance. Then, the relevant cells are cross checked with the selected cells from LSTM_2 layer to see if they have an above average weight in LSTM_2 layer as well. If so, the word will be selected as a word significance.

Chapter 5 – Evaluation and Results

5.1 Evaluation of the model

After training the proposed model for 12 epochs the model becomes saturated and reached an overall accuracy of 0.629. Figure 10 shows the accuracy of the model during training.

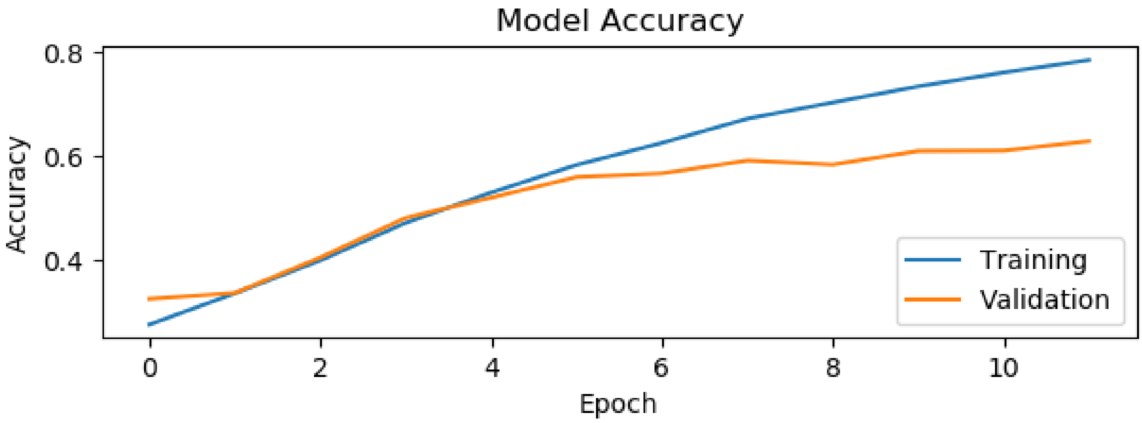

Figure 10: Accuracy of the model

There are four types of outcomes that can occur as this is a classification based model. Those are true positives, true negatives, false positives and false negatives. Based on the outcome, the accuracy, precision and recall of the model can be measured as below.

Accuracy = correct predictions / all predictions = 100/182=0.549

Confusion Matrix of above results is shown below in Table 2.

	Predicted Label									
		US	UK	SL	PK	ML	KE	IN	JP	CN
True Label	US	28	9	0	1	2	5	0	0	1
	UK	4	39	1	0	0	3	1	0	5
	SL	0	0	2	0	0	2	0	0	0
	PK	0	0	0	5	3	3	3	0	2
	ML	0	3	1	0	12	2	0	0	1
	KE	4	2	0	0	1	4	0	0	0
	IN	0	0	1	0	1	1	1	0	1
	JP	2	2	0	0	0	2	0	0	0
	CN	1	0	7	1	0	1	3	0	9

Table 2: Confusion Matrix

True positives = 100

False positives = 82

True negatives = 100

False negatives = 82

Precision = true positives / (true positives + false positives) =100 /182= 0.549

Recall = true positives / (true positives + false negatives) =100 /182 = 0.549

Based on the results, it is evident that even though the model showed 62% an accuracy during training only 54% was achieved.

By analysing the weights of the LSTM layers, set of features were identified that helps in identifying authors' ethno-nationality as well as answer three research questions. Stylistic features used for comparison were selected by analysing the features used for comparison in previous researches such as M. Koppel et al. [3] and R. Hickey et al [12].

5.2 Stylistic features between native and non-native authors

Authors from USA and UK were considered as native authors and authors from Sri Lanka, India, Pakistan, Kenya, Malaysia, Japan and China were considered as non-native authors. At a glance the number of unique words with higher weight, found in posts by native authors is larger than the number of unique words found in posts by non-native authors. Eg:- 2399 for 100 native posts, 1449 for 100 non-native posts. This suggests that the vocabulary in posts by native authors are richer than non-native authors. Features identified for native and non-native authors can be seen in Table 3.

Feature	Eg:-	Native	Non-Native
Definite articles	The, a, an	365	210
Single word prepositions	In, on, as, at, from	813	637
Multi word prepositions	Instead of, as well as, according to	11	9
Shorten forms	I've, we've, haven't	60	24
Long forms	Cannot, did not	35	19
Adverbs	Honestly, occasionally	22	11

Table 3: Features identified for native and non-native authors

According to the above table it is evident that the native authors tend to use the definite articles and single word prepositions more than the non-native authors when writing. The higher use of shorten forms in native authors' texts suggests that they use the language more freely and informally as oppose to non-native authors.

5.3 Stylistic features of South Asian authors

Authors from Sri Lanka, India and Pakistan were considered as South Asian authors. Since they belong to the non-native category their text were analysed against the text from other non-native authors by analysing 100 posts from each category. Although the usage of definite articles, shorten and long forms is similar among both Non-South Asians & South Asians, out of the features used for comparison, there is a clear difference in use of single word prepositions as South Asian authors tend to use more of them. This makes South Asian authors similar to native authors. This could be due to the fact that, the countries considered for South Asian category belongs to the Outer Circle of English, where English is taught as a second language in schools which improves the vocabulary and fluency of authors. Features identified for South Asian and Non-South Asian authors can be seen in Table 4.

Feature	Eg:-	Non-South Asian	South Asian
Definite articles	The, a, an	208	204
Single word prepositions	In, on, as, at, from	495	680
Multi word prepositions	Instead of, as well as, according to	7	7
Shorten forms	I've, we've, haven't	24	25
Long forms	Cannot, did not	25	21
Adverbs	Honestly, occasionally	21	15

Table 4: Features identified for South Asian and Non-South Asian authors

5.4 Stylistic features of Sri Lankan authors

To understand the stylistic features of authors from Sri Lanka, text from Sri Lankan authors were analysed against the text from other non-native authors as well as native authors. The comparison was done by analysing 100 posts from each category. Features identified for Sri Lankan authors can be seen in Table 5.

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
Definite articles	The, a, an	365	208	179	162
Single word prepositions	In, on, as, at, from	813	495	557	463
Multi word prepositions	Instead of, as well as, according to	11	7	9	7
Shorten forms	I've, we've, haven't	60	24	25	12
Long forms	Cannot, did not	35	25	19	19
Adverbs	Honestly, occasionally	22	21	19	5

Table 5: Features identified for Sri Lankan authors

According to the above table, Sri Lankan authors share features such as usage of definite articles, multi word prepositions and long forms with other South Asian authors. Sri Lankan authors also show less use of single word prepositions compared to other South Asian Authors. As evident in the table above, Sri Lankan authors also use relatively less amount of shorten forms and adverbs as oppose to native, non-native and other South Asian authors. Lack of shorten forms suggests that Sri Lankans tend to use the language more formally than other South Asian authors.

Chapter 6 – Conclusion and Future Work

In this research, the task of identifying an author's ethno-nationality has been investigated on a set of blog posts written in English. As compared to previous works done in similar areas such as native language identification, this research takes a new approach with the use of neural networks.

The final model is capable of identifying the trained group of ethnicities with an accuracy of 63%. It has also been shown that the model can identify native authors and non-native authors with an accuracy of 71% and 58% respectively. Since the task of identifying the ethno-nationality and identifying the stylistic features of Sri Lankan authors has not been undertaken previously, there are no previous results to compare this with.

Apart from analysing the feasibility of using neural networks for this type of research, this research tackles the problem of extracting features from LSTM models as well. By using an innovative method of analysing the weights assigned for words by LSTM cells during predictions we managed to identify the stylistic features. Frequent occurrence of selected words with higher weights, within posts from one country proves the hypothesis that authors from the same ethno-nationality will share stylistic features such as choice of words, subject-verb-object order, use of prepositions etc. in their written text. By grouping posts from each country into groups such as native, non-native, South Asian and Sri Lankan gave the opportunity to analyse the stylistic features of authors from each group.

However the results of this research show that there is room for improvement. SVMs are capable of giving better results with less data. Neural networks perform better with more data. Otherwise they tend to be over fitting. This was one of the major challenges faced during this research. The results as well as the accuracy of the model developed could be improved with more data. Currently the model can predict authors from only 9 countries. With more data, the model could be improved to identify authors from more countries with higher accuracy. This could be done as an enhancement to the current model in the future.

Since gathering data is expensive and time consuming, a possible solution would be to apply techniques such as transfer learning. It involves the use of pre trained models to simplify our tasks. Finding such model is difficult as there are not many pre trained models that performs tasks which are useful for this research. Developing a pre trained model is a task that can be handled by the future researchers.

Another fact that was identified during the literature review was how authors from different parts of the same country writes English differently. As a part of future work, one can analyse the differentiation of the same language within a country.

References

[1] "History of English", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/History_of_English. [Accessed: 05- Sep- 2018].

[2] E. Kochmar, "Identification of a Writer's Native Language by Error Analysis", Master of Philosophy in Advanced Computer Science, University of Cambridge, 2011.

[3] M. Koppel, J. Schler and K. Zigdon, "Determining an Author's Native Language by Mining a Text for Errors", in 11th ACM SIGKDD international conference on Knowledge discovery in data mining, 2005.

[4] N. Honna, "English as a Multicultural Language in Asia and Intercultural Literacy", Intercultural Communication Studies, vol. 14, no. 2, 2005.

[5] S. Biesenbach-Lucas, "Students Writing Emails To Faculty: An Examination Of E-politeness Among Native And Non-native Speakers Of English", Language Learning & Technology, vol. 11, no. 2, pp. 59-81, 2007.

[6] A. Prasithrathsint, "Linguistic markers and stylistic attributes of hedging in English academic papers written by native and non-native speakers of English", MANUSYA: Journal of Humanities, vol. 18, pp. 1-22, 2015.

[7] P. Duppenthaler, "A Comparison of Essays Written by Native and Non native Speakers of English on the Topic Kokusai Shakai (International Society)", 2006.

[8] S. Th. Gries, B. Heller and T. Bernaisch, "Epicenters in South and South-East Asian Englishes: Empirical Perspectives", Hong Kong, 2016.

[9] H. F. Schiffman, "Bilingualism in South Asia: Friend or Foe?", International Symposium on Bilingualism (ISBL4), 2003.

[10] B. B. Kachru, "English as an Asian Language", Links & Letters, vol. 5, pp. 89-108, 1998.

[11] T. McArthur, "English as an Asian Language", ABD, vol. 33, no. 2, 2002.

[12] R. Hickey and C. Essen, "English in Asia - The emergence of new varieties", University of Duisburg-Essen, 2011.

[13] "World Englishes", *En.wikipedia.org*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/World_Englishes. [Accessed: 05- Sep- 2018].

[14] M. Koppel, S. Argamon and A. Shimoni, "Automatically Categorizing Written Texts by Author Gender", *Literary and Linguistic Computing*, 2002.

[15] D. Estival, T. Gaustad, B. Hutchinson, S. Pham and W. Radford, "Author Profiling for English and Arabic Emails", Hdl.handle.net, 2008. [Online]. Available: http://hdl.handle.net/2123/5839. [Accessed: 22- Oct- 2018].

[16] K. Santosh, R. Bansal, M. Shekhar and V. Varma, "Author Profiling: Predicting Age and Gender from Blogs", 2013.

[17] Y. Miura, T. Taniguchi, M. Taniguchi and T. Ohkuma, "Author Profiling with Word+Character Neural Attention Network", in *Notebook for PAN at CLEF 2017*, 2017.

[18] "English language", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/English_language. [Accessed: 16- July 2018].

[19] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", Machine Learning: ECML-98, pp. 137-142, 1998.

[20] N. Donges, "The Random Forest Algorithm", machinelearning-blog.com, 2018. [Online]. Available: https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/. [Accessed: 22- Oct- 2018].

[21] M. Lui and P. Cook, "Classifying English Documents by National Dialect", in Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013), 2013.

[22] D. Estival, T. Gaustad, S. Pham, W. Radford and B. Hutchinson, "Author Profiling for English Emails", in 10th Conference of the Pacific Association for Computational Linguistics, University of Melbourne, Melbourne, Australia, 2007, pp. 263-272.

[23] "What is Text Analysis", Ontotext, 2018. [Online]. Available: https://www.ontotext.com/knowledgehub/fundamentals/text-analysis/. [Accessed: 15- Nov- 2018].

[24] "Artificial Neural Networks as Models of Neural Information Processing", Frontiers, 2018. [Online]. Available: https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing. [Accessed: 17- Nov- 2018].

[25] "What is an artificial neural network? Here's everything you need to know | Digital Trends", Digital Trends, 2018. [Online]. Available: https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/. [Accessed: 17- Nov- 2018].

[26] "A Beginner's Guide to LSTMs and Recurrent Neural Networks", Skymind, 2018. [Online]. Available: https://skymind.ai/wiki/lstm. [Accessed: 01- Nov- 2018].

[27] "Introduction to Neural Networks, Advantages and Applications", Towards Data Science, 2018. [Online]. Available: https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207. [Accessed: 18- Nov- 2018].

[28] "Keras LSTM tutorial - How to easily build a powerful deep learning language model - Adventures in Machine Learning", Adventures in Machine Learning, 2018. [Online]. Available: http://adventuresinmachinelearning.com/keras-lstm-tutorial/. [Accessed: 20- Nov- 2018].

[29] J. Brownlee, "How to Diagnose Overfitting and Underfitting of LSTM Models", Machine Learning Mastery, 2017. [Online]. Available: https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/. [Accessed: 21- Nov- 2018].

[30] J. Jordan, "Evaluating a machine learning model.", Jeremy Jordan, 2017. [Online]. Available: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/. [Accessed: 21- Nov- 2018].

[31] "Python Data Analysis Library — pandas: Python Data Analysis Library", Pandas.pydata.org, 2019. [Online]. Available: https://pandas.pydata.org/index.html. [Accessed: 01- May- 2019].

[32] "NumPy — NumPy", Numpy.org, 2019. [Online]. Available: https://www.numpy.org/. [Accessed: 01- May- 2019].

[33] "os — Miscellaneous operating system interfaces — Python 3.7.3 documentation", Docs.python.org, 2019. [Online]. Available: https://docs.python.org/3/library/os.html. [Accessed: 01- May- 2019].

[34] "TensorFlow", TensorFlow, 2019. [Online]. Available: https://www.tensorflow.org/. [Accessed: 01- May- 2019].

[35] "Natural Language Toolkit — NLTK 3.4.1 documentation", Nltk.org, 2019. [Online]. Available: https://www.nltk.org/. [Accessed: 01- May- 2019].

[36] "re — Regular expression operations — Python 3.7.3 documentation", Docs.python.org, 2019. [Online]. Available: https://docs.python.org/3/library/re.html. [Accessed: 01- May- 2019].

[37] "15.3. time — Time access and conversions — Python 2.7.16 documentation", Docs.python.org, 2019. [Online]. Available: https://docs.python.org/2/library/time.html. [Accessed: 01- May- 2019].

[38] "sklearn.model_selection.train_test_split — scikit-learn 0.20.3 documentation", Scikit-learn.org, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. [Accessed: 01- May- 2019].

[39] "Text Preprocessing - Keras Documentation", Keras.io, 2019. [Online]. Available: https://keras.io/preprocessing/text/. [Accessed: 01- May- 2019].

[40] "Sequence Preprocessing - Keras Documentation", Keras.io, 2019. [Online]. Available: https://keras.io/preprocessing/sequence/. [Accessed: 01- May- 2019].

[41] "8.3. collections — High-performance container datatypes — Python 2.7.16 documentation", Docs.python.org, 2019. [Online]. Available: https://docs.python.org/2/library/collections.html#collections.namedtuple. [Accessed: 01- May- 2019].

[43] "Embeddings | TensorFlow Core | TensorFlow", TensorFlow, 2019. [Online]. Available: https://www.tensorflow.org/guide/embedding. [Accessed: 01- May- 2019].

[44] "Core Layers - Keras Documentation", Faroit.com, 2019. [Online]. Available: http://faroit.com/keras-docs/1.2.0/layers/core/#spatialdropout1d. [Accessed: 01- May- 2019].

[45] "Understanding RMSprop — faster neural network learning", Towards Data Science, 2019. [Online]. Available: https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a. [Accessed: 01- May- 2019].

[46] "Kulbear/deep-learning-nano-foundation", GitHub, 2019. [Online]. Available: https://github.com/Kulbear/deep-learning-nano-foundation/wiki/ReLU-and-Softmax-Activation-Functions. [Accessed: 01- May- 2019].

[47] D. Abeygunawardana, "Palaweni wadasatahana", Vb6-in-sinhala.blogspot.com, 2019. [Online]. Available: http://vb6-in-sinhala.blogspot.com/2009/03/palaweni-wadasatahana.html. [Accessed: 03- Nov- 2018].

[48] "Profetas de la vida", Saccvi.blogspot.com, 2019. [Online]. Available: http://saccvi.blogspot.com/2019/04/profetas-de-la-vida.html. [Accessed: 01- May- 2019].

[49] "Asking Too Much", Caringisaverb.blogspot.com, 2019. [Online]. Available: https://caringisaverb.blogspot.com/2017/06/asking-too-much.html. [Accessed: 02- Apr- 2019].

[50] "Common Agricultural Policy: Gove warns of Brexit farming woes", Commonagpolicy.blogspot.com, 2019. [Online]. Available: http://commonagpolicy.blogspot.com/2019/01/gove-warns-of-brexit-farming-woes.html. [Accessed: 02- Apr- 2019].

[51] "How to Remove Acne with Radish", Anihealthtip7.blogspot.com, 2019. [Online]. Available: https://anihealthtip7.blogspot.com/2019/02/how-to-remove-acne-with-radish.html. [Accessed: 02- Apr- 2019].

[52] B. Brown, "Good Times with Wildflowers - Brenda", Countryviewcrafts.blogspot.com, 2019. [Online]. Available: http://countryviewcrafts.blogspot.com/2019/04/good-times-with-wildflowers-brenda.html. [Accessed: 01- May- 2019].

[53] J. Hashmi, "Website Submission Directories", 9oilpaintings.blogspot.com, 2019. [Online]. Available: https://9oilpaintings.blogspot.com/2017/08/website-submission-directories.html. [Accessed: 02- Apr- 2019].

[54] "Introduction | Blogger | Google Developers", Google Developers, 2019. [Online]. Available: https://developers.google.com/blogger. [Accessed: 02- Jan- 2019].

Appendix A - Sample Data

Following are some sample posts from each country.

Country	Post
UK	The text of Michael Gove's speech to the Oxford Farming Conference: Defra Secretary He comments, 'I cannot, here, entirely pre-empt the outcome of the Governmentâ€™s Spending Review.' Indeed, but it is of crucial importance and the Treasury has a long held suspicion of farming subsidies. Gove claims, 'Embracing change, supporting reform is the key to unlocking the Treasuryâ€™s special box.' The Secretary of State admitted, 'Itâ€™s a grim but inescapable fact that in the event of a no-deal Brexit, the effective tariffs on beef and sheep meat would be above 40% - in some cases well above that. While exchange rates might take some of the strain, the costs imposed by new tariffs would undoubtedly exceed any adjustment in the currency markets.' In addition, 'The combination of significant tariffs when none exist now, friction and checks at the border when none exist now and requirements to re-route or pay more for transport when current arrangements are frictionless, will all add to costs for producers. As will new labelling requirements, potential delays in the recognition of organic products, potentially reduced labour flows and the need to provide export health certificates for the EU market which are not needed now.' 'Nobody can be blithe or blasÃ© about the real impact on food producers of leaving without a deal.'[50]
US	When I ask... who cares? am I asking too much? staring at the silence in your eyes All you show are fears when I reach out to touch you seem to be comfortable with lies you run from honest love and defend your pretenses truly unconditional love just fires up your defenses and the open giving heart is used, abused, or taken for granted except in fairy tales where we pretend love is enchanted When I ask... who cares? is it too many questions? tacit acceptance of mutual disguise All you show are fears in your passive rejections protected by caution painted wise you run from honest love... maybe another chorus and refrain will come before the question comes around again maybe another reason and a rhyme will excuse the wasted love and waste of time ... .. .[49]
PK	How to Remove Acne with Radish Step by step instructions to evacuate skin inflammation with radish - Imagine a scenario in which you have skin break out face. It would be aggravating appearance and dispose of fearlessness. Numerous items and magnificence medicines are offered to take care of the issue of skin break out, once in a while much a few people the wrong pick items and treatment. What's more, subsequently considerably more skin break out. Numerous individuals don't have a clue about the normal materials, for example, radishes can lessen skin break out without reactions. For instance, the reactions from the utilization of synthetic substances that can cause the skin ends up aroused. Radish is a vegetable which has numerous advantages for wellbeing and magnificence. Promptly accessible and reasonable, this plant is frequently used to decrease At that point wash utilizing clean water. Covers radishes can likewise be blended with regular fixings, for example, olive oil and nectar to the face winds up smooth and soggy. Veils radishes ought to be utilized when skin inflammation dried.

Country	Post
	This is to abstain from consuming in the face. Covers radishes ok for the skin since it is a characteristic material with the substance nutrient C is high. To complete the most routinely consistently before rest until the skin break out vanishes. The recuperating procedure of skin inflammation with radishes veils dangling from the numerous and extreme skin inflammation or not. Who have extreme irritation of skin inflammation ought to be dealt with quickly by skin expert specialist. Anyway forestall skin break out is superior to fix after the skin break out that emerge. In this manner, abstain from eating high fat sustenances since it would make oil generation in the body increments with the goal that the skin break out effectively emerge. As often as possible beverage water each day is additionally a characteristic method to avert skin break out. Ideally the article How to evacuate skin break out with radish helpful to you. Positive reasoning and solid living propensity. [51]
UK	I have soooo fallen in love the new wildflowers from Tim Holtz I decided to make another card with them. This one is rather bold and striking, definitely not my usual colour palette. Process steps Spritz the back of the doily stencil and sprinkle over sunset beach infusions. Lay over the card panel and rub over with a dry piece of kitchen paper. Remove, heat dry and edge with age mahogany distress ink. Gather together some ephemera pieces to make a little collage .... and adhere the flowers and sentiment over the top. If I were to make another of these now it would take less than half an hour but the thinking time and die-cutting time (I cut 6 different colours of flower!!!) took me several hours to get it right. BUT .... I actually really like it now I've got it done. I hope you do to. Have a lovely April. hugs Brenda xxx Bumblebees and Butterflies [52]
IN	Hello Friends - I have created this post for my fellow Portrait Artists who are searching high page rank directories lists, after receiving lots of request from my artist friends and also from my blog's visitors. I hope you will like this post... please browse very high quality PR website submission directories where you can submit your website URL free of cost and promote your painting business online by submitting your website there, i have checked out personally these sites and all are 100% working. What Is Special In This Kind Of Directories I know and also well knowledge people knows it very well that newbie doesn't have any idea about this kind of submission directories but seasoned internet surfer knows about it very well. I explain you... for example:- you have a website or blog and you have posted many types of articles or something like that but google, yahoo, bing, etc did not indexed it to their search engines so there will be no visitors for your website or blog if you will not have visitors for your site so how people see your website so you need a best and high PR directories lists to submit your website or blog URL there, it is a true that effective links building works for website's owners very well therefore i am bringing this lists for you. There is no need to sign up just open any of the below mentioned URL and choose submit site tab then follow instruction like: select category, title of post, description, keywords etc. and submit website link button that's it. Directories can take times for approval some may take 24 to 48 hours and some 3 to 4 months and some can be instant approved, actually all the submitted URL are manually approved. but if you have financial freedom and want instant approval so can use premium services features too. Keep It In Mind Before Submitting Create unique and keyword match title. Description related to your post. Please share your experience with me about

Country	Post
	these kind of directories by posting message in below comments box after this post. List Of Website Submission Directories List Of Websites Who Require Reciprocal Links Viet Sites Directory Royallinkup Free Website Link Directory DrTest.net Alabama Index BrestLinks Directory Many more links are coming soon please visit this page as soon as possible for latest updates. If you want to get more traffic to your website or blog then try this website also urlLink i hope it will boost your website fast and quickly [53].

Table 6: Sample data

Appendix B - Source Code

Code of the LSTM Neural Network

Model.py

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.models import Model, Sequential
from keras.layers import Input, Dense, Embedding, LSTM, SpatialDropout1D
from sklearn.model_selection import train_test_split
from keras.utils.np_utils import to_categorical
from keras.callbacks import EarlyStopping
from keras.layers import Dropout
import re
from keras.callbacks import TensorBoard
from nltk import word_tokenize
from time import time
from keras.models import model_from_json
import os
import random

df = pd.read_csv('train.csv')
df.head()
df.info()
df.Label.value_counts()

df = df.reset_index(drop=True)
REPLACE_BY_SPACE_RE = re.compile('[/(){}\[\]\|@,;]')
BAD_SYMBOLS_RE = re.compile('[^0-9a-z #+_]')

def clean_text(text):
    text = text.lower() # lowercase text
    text = REPLACE_BY_SPACE_RE.sub(' ', text)
    text = BAD_SYMBOLS_RE.sub('', text)
    return text

df['Post'] = df['Post'].apply(clean_text)
df['Post'] = df['Post'].str.replace('\d+', '')

# The maximum number of words
MAX_NB_WORDS = 179724
# Max number of words in each post.
MAX_SEQUENCE_LENGTH = 200
EMBEDDING_DIM = 200

tokenizer = Tokenizer(num_words=MAX_NB_WORDS, filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',
lower=True)
tokenizer.fit_on_texts(df['Post'].values)
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))
```

```
X = tokenizer.texts_to_sequences(df['Post'].values)
X = pad_sequences(X, maxlen=MAX_SEQUENCE_LENGTH)
print('Shape of data tensor:', X.shape)

Y = pd.get_dummies(df['Label']).values
print('Shape of label tensor:', Y.shape)

X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.10, random_state = 42)
print(X_train.shape,Y_train.shape)
print(X_test.shape,Y_test.shape)
myInput = Input(shape=(MAX_SEQUENCE_LENGTH,), name='input')
x = Embedding(output_dim=EMBEDDING_DIM, input_dim=MAX_NB_WORDS,
input_length=MAX_SEQUENCE_LENGTH)(myInput)
sd = SpatialDropout1D(0.5)(x)
lstm1 = LSTM(100, return_sequences=True, dropout=0.6)(sd)
lstm2 = LSTM(100, dropout=0.6)(lstm1)
predictions = Dense(9, activation='softmax')(lstm2)
model = Model(inputs=myInput, outputs=predictions)

model.compile(loss='categorical_crossentropy',
          optimizer='rmsprop',
          metrics=['accuracy'])
print(model.summary())

epochs = 12
batch_size = 62

tensorboard = TensorBoard(log_dir="logs/{}".format(time()))

history = model.fit(X_train, Y_train, epochs=epochs, batch_size=batch_size, validation_data=(X_test,
Y_test), callbacks=[tensorboard])

accr = model.evaluate(X_test,Y_test)
print('Test set\n  Loss: {:0.3f}\n  Accuracy: {:0.3f}'.format(accr[0],accr[1]))

plt.figure(1)

# summarize history for accuracy
plt.subplot(211)
plt.plot(history.history['acc'])
plt.plot(history.history['val_acc'])
plt.title('Model Accuracy')
plt.ylabel('Accuracy')
plt.xlabel('Epoch')
plt.legend(['Training', 'Validation'], loc='lower right')

# summarize history for loss
plt.subplot(212)
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('Model Loss')
plt.ylabel('Loss')
plt.xlabel('Epoch')
plt.legend(['Training', 'Validation'], loc='upper right')

plt.tight_layout()
plt.show()
```

```
# serialize model to JSON
model_json = model.to_json()
with open("model.json", "w") as json_file:
    json_file.write(model_json)
# serialize weights to HDF5
model.save_weights("model.h5")
print("Saved model to disk")
```

load_model.py

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.models import Model, Sequential
from keras.layers import Input, Dense, Embedding, LSTM, SpatialDropout1D
from sklearn.model_selection import train_test_split
from keras.utils.np_utils import to_categorical
from keras.callbacks import EarlyStopping
from keras.layers import Dropout
import re
from keras.callbacks import TensorBoard
from keras.models import model_from_json
from nltk import word_tokenize
from time import time
import os
import random


df = pd.read_csv(‘train.csv')
df.head()
df.info()
df.Label.value_counts()


df = df.reset_index(drop=True)
REPLACE_BY_SPACE_RE = re.compile('[/(){}\[\]\|@,;]')
BAD_SYMBOLS_RE = re.compile('[^0-9a-z #+_]')
#STOPWORDS = set(stopwords.words('english'))

def clean_text(text):
    text = text.lower() # lowercase text
    text = REPLACE_BY_SPACE_RE.sub(' ', text)
    text = BAD_SYMBOLS_RE.sub('', text)
    return text
df['Post'] = df['Post'].apply(clean_text)
df['Post'] = df['Post'].str.replace('\d+', '')

# The maximum number of words
MAX_NB_WORDS = 179724
# Max number of words in each post
MAX_SEQUENCE_LENGTH = 200
EMBEDDING_DIM = 200
```

```
tokenizer = Tokenizer(num_words=MAX_NB_WORDS, filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',
lower=True)
tokenizer.fit_on_texts(df['Post'].values)
word_index = tokenizer.word_index
# print('Found %s unique tokens.' % len(word_index))

# load json and create model
json_file = open('model.json', 'r')
loaded_model_json = json_file.read()
json_file.close()

### layer 3 ###
model = model_from_json(loaded_model_json)
# load weights into new model
model.load_weights("model.h5")
# print("Loaded model from disk")

model = Model(inputs=model.inputs, outputs=model.layers[3].output)
model.summary()


### layer 4 ###
model1 = model_from_json(loaded_model_json)
# load weights into new model
model1.load_weights("model.h5")
# print("Loaded model from disk")

model1 = Model(inputs=model1.inputs, outputs=model1.layers[4].output)
model1.summary()


### layer 5 ###
model2 = model_from_json(loaded_model_json)
# load weights into new model
model2.load_weights("model.h5")
# print("Loaded model from disk")

model2 = Model(inputs=model2.inputs, outputs=model2.output)
model2.summary()

dx = pd.read_csv('test-all.csv')
dx.head()
dx.info()

ixs = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]

for i in ixs:
   print(i)
   stringx=dx.Post[i-1]

   stringx=clean_text(stringx)
   stringx=stringx.replace('\d+', '')

   new_post1 = []
   new_post1.append(stringx)
   seq = tokenizer.texts_to_sequences(new_post1)
   padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)
```

```
  ###############make predictions#########################
  pred = model.predict(padded) #output of lstm_1
  pred1 = model1.predict(padded) #output of lstm_2
  pred2 = model2.predict(padded)  #output of dense_1

  # get the cells with above average weights in lstm_2
  lstm2_all = [] #all the weights in lstm_2
  maxWeight_lstm2 = [] #max the weights in lstm_2
  maxCells_lstm2 = [] #cells with max the weights in lstm_2

  for a in pred1:
    i=0
    for b in a:
      lstm2_all.append(b)
      if b > np.average(a):
        maxWeight_lstm2.append(b)
        maxCells_lstm2.append(i)
      i += 1

  ### get the cells with above average weights in lstm_1
  lstm1_all = [] #all the weights in lstm_1
  maxWeight_lstm1 = [] #max weight for each word (filtered with lstm_2)
  maxCells_lstm1 = [] #cell with max weight for each word
  maxWord_lstm1 = [] #word token with max weight
  maxWeightAll_lstm1 = [] #max weight for each word
  maxW_lstm1 = []

  ### get the average weight in lstm_1
  for a in pred:
    for b in a:
      maxW_lstm1.append(np.amax(b)) # get max weight of each word (b)

  avg_lstm1 = np.average(maxW_lstm1)

  for a in pred:
    i=0

    for b in a:
      lstm1_all.append(b) # get all weights for one word(b)
      cellNumber = b.argmax() # get cell number of max weight for particular word (b)

      maxWeightAll_lstm1.append(np.amax(b)) # get max weight of each word (b)
      mask = np.isin(cellNumber, maxCells_lstm2) # check whether the selected cell has a higher
weight in lstm_2 also

      if mask == True:
        if np.amax(b) > avg_lstm1:
          maxWord_lstm1.append(i)
          maxWeight_lstm1.append(np.amax(b))
          maxCells_lstm1.append(cellNumber)
      i += 1

  word_list = []
  for x in maxWord_lstm1:
    for t in padded:
      word_list.append(t[x])
```

```
npArray = np.array([word_list])
selectedWords = tokenizer.sequences_to_texts(npArray)
print(selectedWords)

### get the prediction from dense_1
labels = ['CN', 'IN', 'JP', 'KE', 'ML', 'PK', 'SL', 'UK', 'US']
print(pred2, labels[np.argmax(pred2)])
```

Appendix C - Results

Sample high weighted word list

<u>Post 1</u>
cafe greenhouse theres ample seating in the orchard and little garden houses museum natural history museum traffic museum for children hermans vegetarian buffet restaurant with lots of outdoor seating overlooking the harbor theres no longer an elevator but you can walk ride up to the bridge for a great view of the city one of things i like about both of the following off beaten space cafe machine idea what coffee beat

<u>Post 2</u>
collision is a driver left in front of a bicyclist when passing a bicyclist proceed in the same direction slowly and leave at least feet between your car and the cyclist when turning left and a bicyclist is approaching from the opposite direction wait for the rider to pass if turning right and a bicyclist is from behind on the right rider and

Stylistic features between native and non-native authors

Table 7 shows the stylistic features between native and non-native authors.

Feature	Eg:-	Native	Non-Native
Definite articles	the	243	145
	a	110	59
	an	12	6
Total		365	210
Single word prepositions	aboard	0	0
	about	9	0
	above	3	0
	across	3	0
	after	7	8
	against	0	1
	along	12	1
	among	1	3
	around	2	1
	as	29	31
	at	25	17
	before	3	10
	behind	1	1
	below	2	6
	beneath	3	0
	beside	1	0
	between	4	2
	beyond	0	0
	but	42	12

Feature	Eg:-	Native	Non-Native
	by	10	20
	despite	1	1
	down	5	1
	during	5	0
	except	0	0
	failing	0	0
	following	3	0
	for	59	57
	from	29	16
	in	68	122
	inside	0	0
	into	12	4
	like	10	6
	minus	0	0
	near	0	2
	next	0	0
	of	132	120
	off	12	1
	on	52	29
	onto	8	0
	opposite	1	0
	out	8	4
	outside	5	0
	over	11	0
	past	2	2
	plus	1	5
	regarding	0	0
	since	1	3
	than	8	4
	through	5	0
	throughout	0	1
	till	1	10
	to	145	88
	toward	0	0
	towards	0	0
	under	0	1
	underneath	0	0
	unlike	0	1
	until	5	1
	up	21	4
	upon	0	0
	via	1	0
	with	42	36
	within	0	1
	without	3	4

Feature	Eg:-	Native	Non-Native
Total		813	637
Multi word prepositions	according to	1	1
	ahead of	0	0
	along with	1	0
	as for	0	2
	aside from	0	0
	because of	2	0
	close to	1	0
	due to	0	0
	except for	0	0
	far from	0	0
	near to	0	0
	next to	0	0
	out of	2	2
	outside of	0	0
	prior to	0	0
	regardless of	0	0
	instead of	1	2
	as far as	0	0
	as well as	1	1
	in addition to	0	0
	in case of	0	0
	in front of	2	1
	on account of	0	0
	on behalf of	0	0
	on top of	0	0
Total		11	9
Shorten Forms	have->ve, not->nt, will->ll, had->d	60	24
Long forms	cannot, does not, will not	35	19
Adverbs	accidentally	0	0
	always	0	0
	angrily	0	0
	anxiously	0	0
	awkwardly	0	0
	badly	0	0
	blindly	0	0
	boastfully	0	0
	boldly	0	0
	bravely	0	0
	brightly	0	0

Feature	Eg:-	Native	Non-Native
	cheerfully	0	0
	coyly	0	0
	crazily	0	0
	defiantly	0	0
	deftly	0	0
	deliberately	1	0
	devotedly	0	0
	doubtfully	0	0
	dramatically	0	0
	dutifully	0	0
	eagerly	0	0
	elegantly	0	0
	enormously	0	0
	evenly	0	0
	eventually	0	0
	exactly	0	0
	faithfully	0	0
	finally	0	0
	foolishly	0	0
	fortunately	1	0
	frequently	0	0
	gleefully	0	0
	gracefully	0	0
	happily	0	0
	hastily	0	0
	honestly	1	0
	hopelessly	0	0
	hourly	0	0
	hungrily	0	0
	innocently	0	0
	inquisitively	0	0
	irritably	0	0
	jealously	0	0
	justly	0	0
	kindly	0	0
	lazily	0	0
	loosely	1	0
	madly	0	0
	merrily	0	0
	mortally	0	0
	mysteriously	0	0
	nervously	0	0
	never	1	1
	obediently	0	0
	obnoxiously	0	0

Feature	Eg:-	Native	Non-Native
	occasionally	1	0
	often	2	1
	only	8	7
	perfectly	0	0
	politely	0	0
	poorly	0	0
	powerfully	0	0
	promptly	0	0
	quickly	0	0
	rapidly	0	0
	rarely	0	0
	regularly	0	0
	rudely	0	0
	safely	1	0
	seldom	1	0
	selfishly	0	0
	seriously	1	1
	shakily	0	0
	sharply	0	0
	silently	0	0
	slowly	1	1
	solemnly	0	0
	sometimes	1	0
	speedily	0	0
	sternly	0	0
	technically	0	0
	tediously	0	0
	unexpectedly	0	0
	usually	0	0
	victoriously	0	0
	vivaciously	0	0
	warmly	0	0
	wearily	0	0
	weekly	1	0
	wildly	0	0
	yearly	0	0
Total		22	11

Table 7: Stylistic features between native and non-native authors

Stylistic features between non-native authors and South Asian authors

Table 8 shows the stylistic features between non-native authors and South Asian authors.

Feature	Eg:-	Non-Native	South Asian
Definite articles	the	157	137
	a	48	61
	an	3	6
Total		208	204
Single word prepositions	aboard	0	0
	about	4	1
	above	0	0
	across	0	0
	after	11	8
	against	2	0
	along	0	1
	among	1	3
	around	3	0
	as	29	32
	at	12	17
	before	2	11
	behind	1	1
	below	0	12
	beneath	0	0
	beside	1	0
	between	3	1
	beyond	0	0
	but	22	10
	by	17	20
	despite	0	1
	down	5	6
	during	1	1
	except	0	0
	failing	0	0
	following	0	0
	for	26	57
	from	12	13
	in	66	124
	inside	1	3
	into	8	4
	like	4	9
	minus	0	0
	near	0	2
	next	0	0
	of	69	138
	off	1	0
	on	19	36
	onto	0	0
	opposite	0	0
	out	11	9

Feature	Eg:-	Non-Native	South Asian
	outside	1	0
	over	1	0
	past	3	5
	plus	5	1
	regarding	0	0
	since	3	1
	than	2	5
	through	3	0
	throughout	1	1
	till	10	4
	to	85	88
	toward	0	0
	towards	0	0
	under	2	0
	underneath	1	0
	unlike	2	0
	until	1	0
	up	11	10
	upon	0	0
	via	0	0
	with	28	40
	within	0	2
	without	5	3
Total		495	680
Multi word prepositions	according to	2	2
	ahead of	0	0
	along with	0	0
	as for	1	1
	aside from	0	0
	because of	1	1
	close to	0	0
	due to	0	0
	except for	0	0
	far from	0	0
	near to	0	0
	next to	0	0
	out of	1	1
	outside of	0	0
	prior to	0	0
	regardless of	0	0
	instead of	0	0
	as far as	0	0
	as well as	1	1
	in addition to	0	0
	in case of	0	0

Feature	Eg:-	Non-Native	South Asian
	in front of	1	1
	on account of	0	0
	on behalf of	0	0
	on top of	0	0
Total		7	7
Shorten Forms	have->ve, not->nt, will->ll, had->d	24	25
Long forms	cannot, does not, will not	25	21
Adverbs	accidentally	0	0
	always	2	1
	angrily	0	0
	anxiously	0	0
	awkwardly	0	0
	badly	0	0
	blindly	0	0
	boastfully	0	0
	boldly	0	0
	bravely	0	0
	brightly	0	0
	cheerfully	0	0
	coyly	0	0
	crazily	0	0
	defiantly	0	0
	deftly	0	0
	deliberately	0	0
	devotedly	0	0
	doubtfully	0	0
	dramatically	0	0
	dutifully	1	0
	eagerly	0	0
	elegantly	0	0
	enormously	0	0
	evenly	0	0
	eventually	0	0
	exactly	0	0
	faithfully	0	0
	finally	1	0
	foolishly	0	0
	fortunately	0	0
	frequently	1	0
	gleefully	0	0

Feature	Eg:-	Non-Native	South Asian
	gracefully	0	0
	happily	0	0
	hastily	0	0
	honestly	0	0
	hopelessly	0	0
	hourly	0	0
	hungrily	0	0
	innocently	0	0
	inquisitively	0	0
	irritably	0	0
	jealously	0	0
	justly	0	0
	kindly	0	0
	lazily	0	0
	loosely	0	0
	madly	0	0
	merrily	0	0
	mortally	0	0
	mysteriously	0	0
	nervously	0	0
	never	3	4
	obediently	0	0
	obnoxiously	0	0
	occasionally	0	0
	often	1	0
	only	7	8
	perfectly	2	0
	politely	0	0
	poorly	0	1
	powerfully	0	0
	promptly	0	0
	quickly	0	0
	rapidly	0	0
	rarely	0	0
	regularly	0	0
	rudely	0	0
	safely	0	0
	seldom	0	0
	selfishly	0	0
	seriously	0	1
	shakily	0	0
	sharply	0	0
	silently	0	0
	slowly	1	0
	solemnly	0	0

Feature	Eg:-	Non-Native	South Asian
	sometimes	2	0
	speedily	0	0
	sternly	0	0
	technically	0	0
	tediously	0	0
	unexpectedly	0	0
	usually	0	0
	victoriously	0	0
	vivaciously	0	0
	warmly	0	0
	wearily	0	0
	weekly	0	0
	wildly	0	0
	yearly	0	0
Total		21	15

Table 8: Stylistic features between non-native authors and South Asian authors

Stylistic features of Sri Lankan authors

Table 9 shows the Stylistic features of Sri Lankan authors.

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
Definite articles	the	243	157	120	101
	a	110	48	56	50
	an	12	3	3	11
Total		365	208	179	162
Single word prepositions	aboard	0	0	0	0
	about	9	4	0	1
	above	3	0	0	3
	across	3	0	1	0
	after	7	11	18	9
	against	0	2	0	0
	along	12	0	0	0
	among	1	1	4	2
	around	2	3	0	0
	as	29	29	19	28
	at	25	12	11	11
	before	3	2	4	2
	behind	1	1	1	0
	below	2	0	13	3
	beneath	3	0	0	0
	beside	1	1	0	0
	between	4	3	1	1

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
	beyond	0	0	0	0
	but	42	22	9	6
	by	10	17	15	14
	despite	1	0	1	0
	down	5	5	0	0
	during	5	1	2	3
	except	0	0	0	0
	failing	0	0	0	0
	following	3	0	0	2
	for	59	26	46	31
	from	29	12	15	9
	in	68	66	100	87
	inside	0	1	3	5
	into	12	8	1	1
	like	10	4	14	1
	minus	0	0	0	0
	near	0	0	2	0
	next	0	0	0	0
	of	132	69	128	87
	off	12	1	0	1
	on	52	19	26	30
	onto	8	0	0	0
	opposite	1	0	0	0
	out	8	11	2	2
	outside	5	1	0	0
	over	11	1	1	1
	past	2	3	5	2
	plus	1	5	9	0
	regarding	0	0	0	3
	since	1	3	2	0
	than	8	2	9	3
	through	5	3	0	2
	throughout	0	1	0	1
	till	1	10	2	0
	to	145	85	60	84
	toward	0	0	0	1
	towards	0	0	0	0
	under	0	2	0	0
	underneath	0	1	0	0
	unlike	0	2	0	0
	until	5	1	0	1
	up	21	11	2	0
	upon	0	0	0	0
	via	1	0	1	0

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
	with	42	28	26	24
	within	0	0	2	0
	without	3	5	2	2
Total		813	495	557	463
Multi word prepositions	according to	1	2	3	2
	ahead of	0	0	0	0
	along with	1	0	0	0
	as for	0	1	2	1
	aside from	0	0	0	0
	because of	2	1	1	0
	close to	1	0	0	0
	due to	0	0	0	1
	except for	0	0	0	0
	far from	0	0	0	0
	near to	0	0	0	0
	next to	0	0	0	0
	out of	2	1	1	0
	outside of	0	0	0	0
	prior to	0	0	0	0
	regardless of	0	0	0	0
	instead of	1	0	0	2
	as far as	0	0	0	0
	as well as	1	1	1	1
	in addition to	0	0	0	0
	in case of	0	0	0	0
	in front of	2	1	1	0
	on account of	0	0	0	0
	on behalf of	0	0	0	0
	on top of	0	0	0	0
Total		11	7	9	7
Shorten Forms	have->ve, not->nt, will->ll, had->d	60	24	25	12
Long forms	cannot, does not, will not	35	25	19	19
Adverbs	accidentally	0	0	0	0
	always	0	2	5	1
	angrily	0	0	0	0
	anxiously	0	0	0	0
	awkwardly	0	0	0	0
	badly	0	0	0	0
	blindly	0	0	0	0

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
	boastfully	0	0	0	0
	boldly	0	0	0	0
	bravely	0	0	0	0
	brightly	0	0	0	0
	cheerfully	0	0	0	0
	coyly	0	0	0	0
	crazily	0	0	0	0
	defiantly	0	0	0	0
	deftly	0	0	0	0
	deliberately	1	0	0	0
	devotedly	0	0	0	0
	doubtfully	0	0	0	0
	dramatically	0	0	0	0
	dutifully	0	1	0	0
	eagerly	0	0	0	0
	elegantly	0	0	0	0
	enormously	0	0	0	0
	evenly	0	0	0	0
	eventually	0	0	1	0
	exactly	0	0	0	0
	faithfully	0	0	0	0
	finally	0	1	0	0
	foolishly	0	0	0	0
	fortunately	1	0	0	0
	frequently	0	1	0	0
	gleefully	0	0	0	0
	gracefully	0	0	0	0
	happily	0	0	0	0
	hastily	0	0	0	0
	honestly	1	0	0	0
	hopelessly	0	0	0	0
	hourly	0	0	0	0
	hungrily	0	0	0	0
	innocently	0	0	0	0
	inquisitively	0	0	0	0
	irritably	0	0	0	0
	jealously	0	0	0	0
	justly	0	0	0	0
	kindly	0	0	0	0
	lazily	0	0	0	1
	loosely	1	0	0	0
	madly	0	0	0	0
	merrily	0	0	0	0
	mortally	0	0	0	0

Feature	Eg:-	Native	Non-South Asian	South Asian	Sri Lankan
	mysteriously	0	0	0	0
	nervously	0	0	0	0
	never	1	3	5	0
	obediently	0	0	0	0
	obnoxiously	0	0	0	0
	occasionally	1	0	0	0
	often	2	1	0	0
	only	8	7	6	2
	perfectly	0	2	0	0
	politely	0	0	0	0
	poorly	0	0	1	0
	powerfully	0	0	0	0
	promptly	0	0	0	0
	quickly	0	0	0	0
	rapidly	0	0	0	0
	rarely	0	0	0	0
	regularly	0	0	0	0
	rudely	0	0	0	0
	safely	1	0	0	1
	seldom	1	0	0	0
	selfishly	0	0	0	0
	seriously	1	0	1	0
	shakily	0	0	0	0
	sharply	0	0	0	0
	silently	0	0	0	0
	slowly	1	1	0	0
	solemnly	0	0	0	0
	sometimes	1	2	0	0
	speedily	0	0	0	0
	sternly	0	0	0	0
	technically	0	0	0	0
	tediously	0	0	0	0
	unexpectedly	0	0	0	0
	usually	0	0	0	0
	victoriously	0	0	0	0
	vivaciously	0	0	0	0
	warmly	0	0	0	0
	wearily	0	0	0	0
	weekly	1	0	0	0
	wildly	0	0	0	0
	yearly	0	0	0	0
Total		22	21	19	5

Table 9: Stylistic features of Sri Lankan authors