Language Detection in Sinhala-English Code-mixed Data

Ian Smith
Computer Science Engineering
University of Moratuwa, Sri Lanka
royian.18@cse.mrt.ac.lk

Uthayasanker Thayasivam
Computer Science Engineering
University of Moratuwa, Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Abstract*— Language identification in text data has become a trending topic due to multiple language usage on the internet and it becomes a difficult task when it comes to bilingual and multilingual communication data processing. Accordingly, this study introduces a methodology to detect Sinhala and English words in code-mixed data and this is the first research done on such scenario at the time of this paper is written. In addition to that, the data set which is used for this research was newly built and published for similar research users. Even though there are well known models to identify Singlish Unicode characters which is a straightforward study; there are no proper language detection models to detect Sinhala words in a sentence which contains English words (code-mixed data). Therefore, this paper presents a language detection model with XGB classifier with 92.1% accuracy and a CRF model with a F1-score of 0.94 for sequence labeling.**

Keywords—code-mixed, code-switching, Sinhala-English, language detection, social media data

I. INTRODUCTION

To begin with, language identification (LID) is the process of determining the natural language or the language features of a document, part of a document or in a line segment. On the other hand, one of the main characteristics of being a human being is the ability to use complex and sophisticated ideas and thoughts within the communication process using a common language. Accordingly, humans have the ability to identify languages or language features swiftly from a given line segment provided that the text contains languages those are which the human is familiar with. Therefore, it is noteworthy that even the humans cannot identify the languages used in a given document if they are not familiar with the languages used in the text. Hence, the aim of a language detection process is to identify an approach for humans to recognize specific languages used in the given text. Subsequently, there are number of researches those have been conducted to develop specific/general language identification models. Likewise, several data structures, algorithms and models have been developed to model the human ability to identify languages. Furthermore, most of the people were using various encoding methods to present text until the Unicode standard (preceded by ISO 8859, various others) was introduced. Subsequently, language detection becomes comparatively easier with the Unicode characters. Indeed, with these standards and algorithms, the ability to detect the languages based on the encodings in the text is straightforward. However, these simple models can only handle text based on the encoding and they simply rely on the encoding boundaries which fails in most of the occasions such as in social media data and other scenarios where code-mixing occurs.

On the other hand, usage of social media has become a day-to-day activity of human lives as a result of the advancement in technology. Accordingly, social media sites such as Facebook, Twitter and Instagram generate vast amount of data from their services. Moreover, users of social media come from all over the world, each bringing their own backgrounds and cultures into the mix [6]. With such global user base, social media becomes a melting pot of languages used in different manners and for different purposes. This has resulted in creation of an alternate dialect in favors real-time communication such as acronyms and short forms of words that are used in instant messages. For example, people often use expressions such as LOL for laugh out loud, ROFL for rolling on floor laughing, NVM never mind and YOLO for you only live once and common reacts like HAHA for funny reactions. Indeed, these acronyms can be mostly seen in chat communications where users tend to type to express themselves with less characters rather than typing the whole meaning.

Even though English is considered as the native language for web communications, there are significant number of people who use social media with various native languages other than English. However, most of these people do not use Unicode characters to write in their language. Instead of Unicode, most of these people use phonetic typing, frequently inserting English elements in their communication. Accordingly, this has become the latest trend in social media due to the simplicity. Therefore, people express their languages using English characters and they even insert English words mixed up with the native language words. This phenomenon is known as the code-mixing or code-switching interchanging in some researches. Accordingly, in this study, code-mixing is used to identify this phenomenon. Further, the code-mixing can be mostly seen in peer to peer communication, group communication, comments and status updates and so on. On the other hand, people tend to mix Unicode characters with English characters as well and therefore, there can be many variations in each social media content data (SMCD) segment as shown in table 1.

TABLE 1: TYPES OF LANGUAGE VARIATIONS IN SOCIAL MEDIA

Content type	Example
Purely written in English	I'm going to the university today
Purely written in Unicode	අද මම විශ්ව විද්යාලයට යනවා
Written in a language other than English using English characters	Ada mama vishva vidyalayata yanawa
English characters with Unicode	අද මම university යනවා
Code mixing	Ada mama university yanawa

Likewise, people tend to use English alphabetical characters to express their language. Therefore, some new language variations have evolved amongst social media users such as, people mix English with their native language in communication (mostly in writing): Chinglish (Chinese written in English), Hinglish (Hindi written in English) and Singlish (Sinhala written in English). In fact, each of these language variations are expressed using English alphabet and some of the social media users combine English with Unicode characters as well. In Sri Lanka, many marketing campaigns conducted on social media, via emails or text messages use Singlish language features to convey the message to the users effectively. Due to the variety of derived languages used in social media, automatic language identification (ALID) has

become a key research area in natural language processing domain where types of languages and language features are to be detected on a given text. Accordingly, language detection models are expected to classify the language used in a given text or a line segment. This is mainly formed on text written in one language and some sophisticated models are used if the given text contain more than one language (code-mixed data). Further, most language identification models perform well on a single or multiple languages, yet, most of them fail when there are texts with mixed language features. Therefore, even though the language detection has been achieved up to a higher level of accuracy, it is not a fully achieved task in NLP domain.

Likewise, there are many elements to be discovered in language detection in NLP domain and these discoveries become difficult due to rapid changes in human behavior. In fact, it is even more challenging the evaluation of social media data where users communicate with multiple language mixtures. Accordingly, this research focuses on social media data (Facebook within this study) and attempts to develop a model to identify languages in code-mixed data. However, this study focuses only on Sinhala-English code-mixed data and for the best of author's knowledge this is the first study to research and develop such a model in Sinhala-English mixed data at the point of this paper is been developed. Therefore, there are no similar models to be compared with and this study will evaluate multiple techniques based on literature.

II. SIMILAR STUDIES

Although the language detection across documents is considered as a solved task, it is not the same with code-mixing. Nevertheless, there are many researches in language detection in code-mixing predominantly in Indian language context for Hindi and English code-mixing, Bangla and English code-mixing [1] [2] and for Hindi, Bangla and English code-mixing [3]. Accordingly, these studies highlighted that the language identification at word-level is a challenging task which is non-trivial specially in noisy and translated data found in social media platforms.

Accordingly, [2] is one of the leading studies conducted to address the automatic language detection problem in word level for code-mixed data found in social media. This study was conducted for Hindi-English and Bengali-English code-mixed data and the authors have used a dictionary-based classifier as the baseline mode, SVM classifier with 4 kinds of features; weighted character n-grams, dictionary features, minimum edit distance weight and word context information. Accordingly, the best performing model results a high precision more than 90% for Hindi-English data and 87% for Bangla-English data. Yet, the model has shown low recall of 65% and 60% respectively for two language mixtures with an overall F1 score of 76% and 74% respectively.

Further, [1] has introduced a language identification model for Hindi-English code-mixed data with a part-of-speech (POS) annotation system on social media data. Accordingly, the study has used a word-level logistic regression model for model training with 3,201 English words scraped from SMS data and with a separate Hindi corpus of 3,218 words. Even though the given model results a F1 score of 87%, it shows a low recall for Hindi data.

On the other hand, [4] presented a system to identify languages in Bangla-English code-mixed data. Accordingly, the study has used two different datasets, one form FIRE 2013 and the second dataset from Facebook chat history. Similarly, the best performing model uses Bangla and English dictionary, n-gram and the percentage of surrounding words predicted as Bangla using the dictionary as features. Likewise, the proposed model gives a F1 score of 91.5% for Facebook chat dataset and 91.5% for FIRE data set.

A well performing model has been introduced by [3] for language identification in code-mixed data for Indian languages. This study has used multilingual dataset of Hindi-English-Bangla and the authors have collected 2,335 Facebook posts and 9,813 comments from Facebook groups for their experiment. Likewise, the best performing model of the study was a CRF model which was trained with 5 different features named; character n-gram, presence/absence in dictionaries, word length, capitalization and contextual information. Accordingly, the final system shows a 95.76% accuracy. On the other hand, the study has also tried an SVM classifier with the same features and has been able to get an accuracy of 95.52%.

Similarly, [5] introduced a new study to identify languages in code-mixed data with decision tress and SVM classifiers. Likewise, the study was presented using Assamese-English-Hindi code-mixed data for the first time in the domain. Accordingly, the study has collected 4,768 Facebook comments with a total of 20,781 tokens after manual annotation task. The proposed model has used only three features named; word unigrams, prefixed and suffixes and finally the contextual information. On account of the results presented within the study, SVM model outperformed the Decision tree model with an accuracy of 96.01%.

Further, [6] have studied language identification in code-mixed data using a dataset with 30 languages. In this study, 6 features have been used individually as well as with combinations of multiple features. Accordingly, character unigrams, bigrams, trigrams, 4-grams, 5-grams, and the full word have used as features to train their models. Likewise, the study has used CRF model trained with GE (generalized expectation) model, HMM (Hidden Markov Model) trained with EM (Expectation Maximization) and finally a logistic regression model trained with GE. Finally, each model has been compared with Naïve Bayes model as the baseline model. Accordingly, CRF model trained with GE has been able to outperform all the other models with an overall accuracy of 95%.

In contrast, [7] has studied code-mixing data with the aim of language identification using audio data and with their translated text data. Accordingly, the study has used 242,475 words of text in English and Spanish languages. Similarly, the authors have used word n-grams, character n-grams and character prefixes/suffixes as the text features in the study. Moreover, CRM model, logistic regression and a deep neural network model named LSTM has been used in the study. Based on the results, the character level and word level features with the combination of CRF model has been able to outperform the deep neural network model LSTM with an accuracy of 91%.

A well-trained model has been introduced by [8] for Persian and Dari text data. Since the Dari is a low-resourced language, the authors have created a new dataset of 28,000 Dari sentences from an American news website. Like most of the scholars, the authors have used character n-gram and word

n-gram features in their training and SVM classifier has been evaluated with the identified features. Accordingly, an accuracy of 96% has been acquired with the proposed model. On the other hand, the study has also tried out of domain cross-corpus evaluation to test the discriminative models' generalizability, achieving 87% accuracy in classifying 79,000 sentences from the Uppsala Persian Corpus.

The work presented by [9] identified a langue detection model for code-mixed data found in Tweeter platform. Accordingly, the study has used a dataset of 1.1 million tweets in five different European languages collected using tweeter scraping methods. The authors have used weighted n-gram features for language identifications on post level and reported an accuracy of 92.4%. On the other hand, [10] investigated a language identification model at the utterance level on a dataset from one of the largest online communities in The Netherlands for Turkish-Dutch speakers. The study has accommodated dictionary-based language models, logistic regression models and linear-chain CRF in analysis and has been able to reach accuracy of 97.6%, but with a substantially lower accuracy on post level 89.5%, even though 83% of the posts were monolingual.

III. Motivation of the Study

The scope of this study is to evaluate Sinhala-English code-mixing data written in English alphabetical characters rather than Sinhala Unicode data. That is, Unicode characters can be directly identified with their base boundaries and therefore, it is straightforward to identify the language in sentences with Unicode characters. Even though there are many studies conducted on code-mixing data analysis, those models cannot be directly applied to Sinhala-English code-mixing scenario. That is, there are many ambiguous words within Sinhala-English code-mixing. As presented in table 2, even though some of the tokens are present in English language, the meaning of each token is completely different when it comes to Sihala-Singlish code-mixing. On the other hand, Sri Lankan users tend to use some words with short forms like "prens" to represent "friends" and "okkk", "oki", "k" to represent "okay". In addition to that, people also tend to use character "k" or "i" at the end of numbers which will eventually turned to Sinhala token. As an example, "100k" is used to indicate exactly 100 rather than 100,000 in common practice.

TABLE 2: AMBIGUOUS WORDS

Token	Annotation	Annotator justification
royal	Name	Name of a school in Sri Lanka
oke	Sinhala	Sinhala term for "that" in English
maxaa	Sinhala	Form of a complement in Sinhala
okkkk	English	Refers to "Okay" in English even though it is similar to "oke" term in Sinhala
sup	English	Short term used for "Support"
shape	English/Sinhala	Even though this is an English term, sometimes it is used in Sinhala to express terms like "never the less" or a compliment etc.
100k, 6k and 14k etc.	Sinhala	In practice, it refers to 100,000 and 6,000. But here it means "exactly 100 and exactly 6" in Sinhala based on the context
prenzz	English	Short term used English term "Friend"
4i, 5i	Sinhala	Refers to "exactly 4, exactly 5" in Sinhala

IV. Methodology

As stated earlier, this is the first Sinhala-English code-mixing analysis study done by the time of this paper is written, there are no proper datasets to be used. Therefore, a new dataset was created to conduct this study using Facebook data. Further, the study was divided into two parts as code-mixing detection and sequence tagging. Each of the studies have used multiple techniques and compared their performance to select the best performing model.

A. Dataset

A new data set was created by scraping Facebook chat history and publicly available page comments and posts [12]. Following a manual cleanup on the collected data, 7,500 sentences with 40,915 tokens were collected. All the sentences with only Unicode characters, and emoticons were removed to get sentences with English alphabetical characters.

The data was annotated by three undergraduate students who are fluent in Sinhala and English languages with the help of Google sheets. The annotation was done in two phases to annotate sentences with code-mixing and to annotate each word in code-mixed data with the language used by the token. In addition, spelling mistakes are common in any kind of communication and it may occur frequently in social media communication where the users are communicating in near real-time manner and social media users tend to use words in short form in practice which also leads to spelling mistakes within analysis. Therefore, each annotator was asked to avoid all spelling errors or short form of words ("Flm" for "Film", "Tkt" for "Ticket" and "Tnx" for "Thanks" etc.) and annotate the erroneous word with its most probable language.

B. Annotation evaluation

Cohen's Kappa measure is used to measure the annotation accuracy which provides the proportion of agreement beyond that expected by chance. That is, the achieved "beyond chance" agreement as a proportion of the possible "beyond chance" agreement [11].

C. Level 1 annotation

In the first level of annotation, each annotator was requested to annotate each of the sentence with the type of language mix used in the sentence. Accordingly, 5 tags were used in this annotation process as shown in table 3.

Accordingly, a total of 7,500 sentences were annotated in three batches of 1,500, 3,000 and 3,000 sentences per batch. Further, the annotations were carried out as one batch at a time and each sentence was annotated by all three annotators. Furthermore, when assigning a data batch to an annotator, the dataset was shuffled before the assignment to maintain the randomness. Each data batch was given with the instruction table (table 3) to guide annotators.

Once the annotation is finished for a given batch, inter annotator agreement was calculated for the batch. Table 4 shows Cohen's Kappa value for each of the level 1 annotation batch. Final annotation of each sentence was decided based on the total agreement of all three annotators. Likewise, table 5 shows language wise summary of each data batch after their annotation. Even though 7,500 sentences were used for annotations, only 7,080(total count in table 5) sentences got total agreement (same annotation tag from all three annotators) and thus, the final level 1 annotated sentences

selected for level two annotations have 7,080 sentences with the annotations.

TABLE 3: LEVEL 1 ANNOTATION TAGS

Language type	Description	Example
English	Every single word in the text is in English language	Good morning
Singlish	Sinhala words written in English characters	Mama ennam
Sinhala (Unicode)	Sinhala words written in Unicode characters	සුබ උදෑසනක්
Code-mixed	Sinhala written in English (as Singlish), but there are some English words in the sentence	mama ennam film hall ekata
	Unicode characters with Singlish	මම yanawa
	English words with Unicode characters	Dreams පුදුම හිතෙනවා
	English words with Unicode characters and Singlish	පුදුම හිතෙන dream ekak
Unknown	Anything other than the above	

TABLE 4: INTER ANNOTATOR AGREEMENT FOR LEVEL 1 ANNOTATION

Data batch ID	Number of sentences	Cohen's Kappa value
1	1,500	0.806948108
2	3.000	0.878913945
3	3,000	0.936152024

At the end of all three data batches, the inter annotation agreement calculation (Kappa value) was 0.88772595 which shows a high agreement between all three annotators. Table 5 illustrates the overall language mix in the total dataset of 7,500 sentences. Accordingly, Singlish is the dominant language used in the data set. Further, there are considerable amount of code-mixed data as well. Even though there are 476 English sentences, English language usage is negligible when considering the total dataset. On the other hand, there are only 13 sentences with the *unknown* annotation which is also negligible. A manual inspection was done to validate all the sentences with unknown tag, and it was clear that those sentences were either with one or more question marks (??) or a sentence with just a number. Furthermore, the dataset consists of single word sentences as well as which are important to the language detection study.

TABLE 5: OVERALL LANGUAGE USAGE

Language type	Sentence count
Singlish	4,691
Code-mixed	1,900
English	476
Unknown	13

D. Level 2 annotation

Level 2 annotation is the word level annotation of the code-mixed sentences selected from level 1 annotations. Accordingly, all the sentences where all three annotators have annotated as code-mixed were selected to be used in the second annotation phase. Likewise, 1,900 code-mixed sentences were used for word level annotation. Since the dataset is created to be used for a language detection study in Sinhala-English code-mixed data, level 2 annotation was intended to annotate each word based on the language used. However, there were many named entities and some numeric characters in the middle of sentences. Therefore, the second level annotation was carried out with two main tags named *Sinhala* and *English* along with additional tags named *Unknown* and *Name*. Name tag was used to annotate all named entities and Unknown tag will be assigned to all the other cases which do not fall into above categories.

Annotators were assigned with the total data set with the annotation slots for each word in the sentence and the dataset consisted of some acronyms and named entities and the annotators were requested to annotate such words based on the base language of those tokens. The important part in this annotation is that the tag of each word does not depend only in the word itself. That is, the language tag depends on the surrounding words of the sentence (sentences with more than one word). Likewise, there can be many instances where a given token may appear as an English word in this context, but it may be a Sinhala word written using English characters (Singlish). As an example, the word "me" in "me ahanna" sentence, appears as an English word (me, myself). Yet, it is a Sinhala term for English word "hey". That is, the true meaning of the sentence is "hey, listen" in practice. Therefore, annotators had to consider language tags of sounding words in the annotation.

Accordingly, 11,795 tokens were annotated in this annotation phase with their base language. Final annotation of each word was decided based on the majority vote in this annotation level to avoid removal of words in the middle of a sentence (to avoid loss of context clues of each word) if the word fails to achieve a total agreement. Since, there were no annotations without a majority vote, the overall inter annotation agreement Kappa statistics for level two annotation was 1. Likewise, 8,568 Sinhala tokens, 2,824 of English tokens, 350 Name tokens and 53 unknown tokens were identified by the annotation.

In this annotation phase, Sinhala words dominated the code-mixed dataset which is an expected scenario from Sri Lankan users. On the other hand, each Sinhala word mixed with English tokens were contributed to the whole sentence to be a code-mixed sentence in the first level annotation. Further, the domination of Sinhala tokens was also expected because, within the level one annotation results were dominated by Singlish sentences.

E. Code-mixed sentence classification

As the first part of the study, a multiple machine learning models were evaluated to classify distinguished Sinhala-English code from non-code-mixed data. Accordingly, 7,486 sentences with four annotations (Singlish, English, code-mixed and unknown) were used to train each model and BOG, word level TF-IDF, char-n-gram TF-IDF, n-gram TF-IDF, pre-trained word embedding (Genisim Word2Vec) model and a word embedding model trained with the dataset were tried with each machine learning model.

As machine learning models, Naïve Bayes model is used as the baseline model and logistic regression, SVM, Random Forest, XGB, shallow NN, deep NN, LSTM, CNN, recurrent CNN, bidirectional RNN and GRU models were tested with the dataset and optimized individually to select the best performing model.

F. Sequence tagging

Word level sequence tagging was done as the second part of the study. Accordingly, 1,900 Sihala-English code-mixed sentences resulted in 11,795 tokens. The annotations consisted of Sinhala, English, Name and unknown tags. Since the study is identifying each token with the language presented, all tokens annotated as Name are also considered as unknown. CRF, LSTM, SVM, K-nearest neighbor and random forest models were trained and optimized with the data set. Likewise, character n-gram, annotation of three to left side and right side of the word, capitalization and whether the token has a digit or not is used as features.

V. Results

Fig 1 shows the comparison of all the models used for sequence tagging with Naïve Bayes model and fig 2 and fig 3 shows the comparison of non-neural network models and neural network models respectively. Accordingly, it is observed that the character n-gram gives the highest accuracy for most of the models. Further, multiple n-grams were tested and finally bigram was selected to use for all the models based on the accuracy.

When comparing all the models, XGB outperformed all the models with an accuracy of 92.1% with bigram features. On the other hand, it is also observed that all the neural network models were underneath the benchmark model accuracy and even the SVM model did not perform well compared to Naïve Bayes model. In addition, the best performing XGB model underperform than the baseline model with n-gram-TF-IDF features and with BOG features.

Further, LSTM and GRU models performed better than other neural network models and it is also with character n-gram feature. On the other hand, deep neural network model has barely touched the 50% accuracy mark. Yet, other neural network models have been able to perform well compared to deep networks even though they were below the benchmark.

Fig 4, fig 5, fig 6 and fig 7 show the results for sequence annotation study with the identified machine learning models. Based on the results, CRF model outperformed all the other models in precision, recall as well as in F1 score which is a formulation of both precision and recall scores. Further, each model showed a high precision and recall for English tokens than Sinhala and unknown tokens. Accordingly, based on fig 7, CRF model results F1 score of 0.94, precision of 0.95 and a recall of 0.94 on average for all three tokens. Finally, the test output of the CRF model is manually inspected to check the ambiguous word labeling capability of the model and the CRF model has been able to label tricky words such as “100k, 5i and royal” based on the language usage.

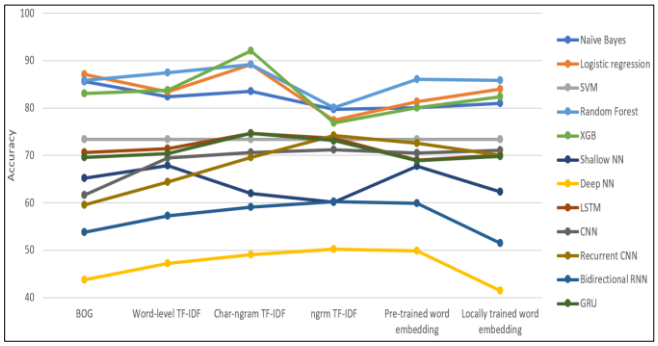

Figure 1: Comparison of all models for code-mixing classification

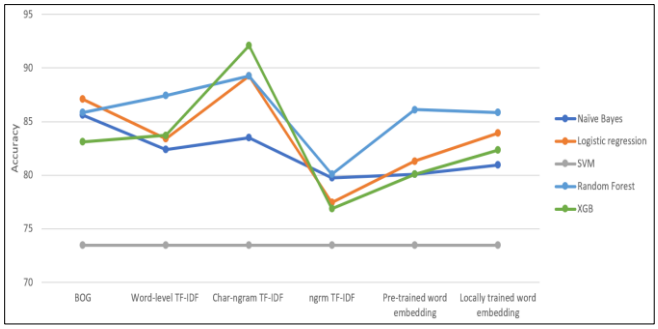

Figure 2: Comparison of non-neural network models for code-mixing classification

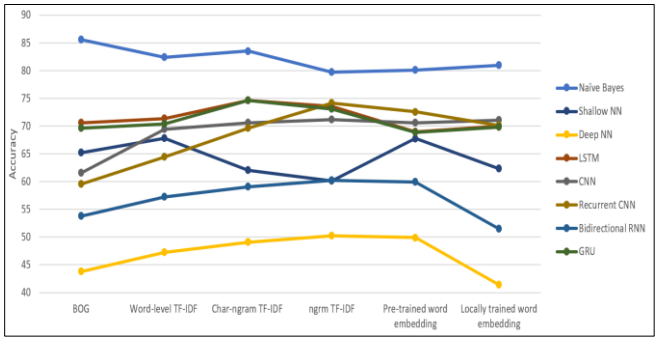

Figure 3: Comparison of neural network models for code-mixing classification

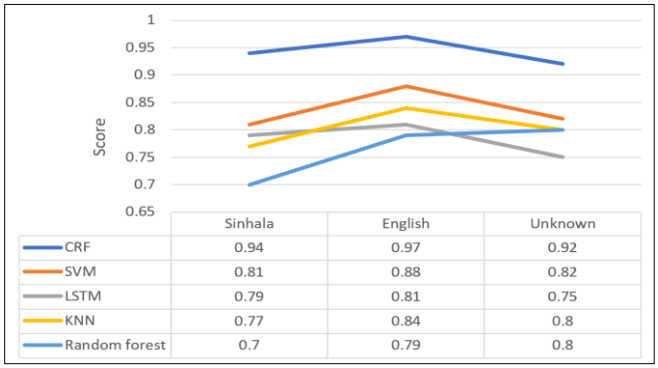

Figure 4: Precision scores

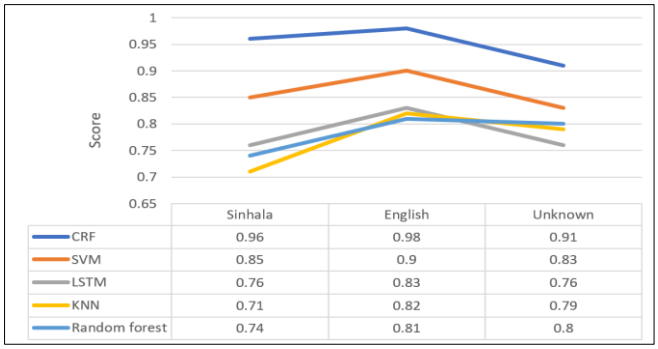

Figure 5: Recall scores

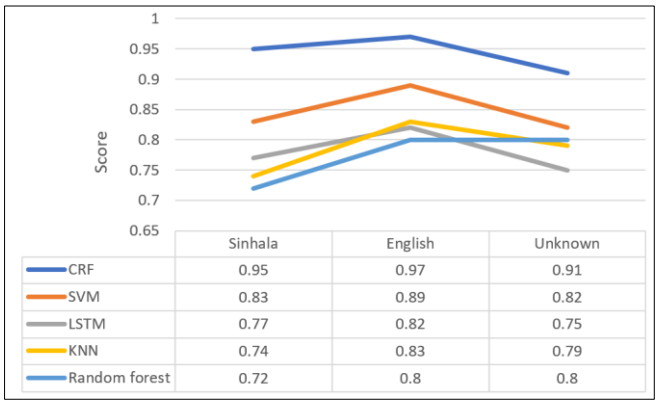

Figure 6: F1 scores

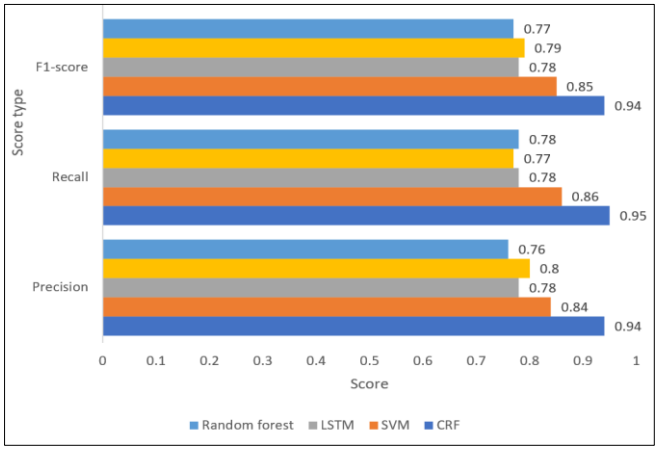

Figure 7: Average scores

VI. DISCUSSION

This study was an attempt to develop a high performing machine learning model to detect language of each token in Sinhala English code-mixed data. Since, this is the first attempt in Sinhala English code-mixed data analysis, a novel data set from Facebook chat history and public posts were created with manual annotation process resulting in high inter annotator agreement. Finally, a well optimized XGB model with an accuracy of 92.1% in code-mixed data classification and a CRF model with an overall F1 score of 0.94 was built in the study after comparing with multiple machine learning models. Even random forest model outperformed XGB in all other cases except for char-n-gram feature. Thus, it was recognized that tree-based models are more suitable for this kind of code-mixed data classification. However, it is deemed that the dataset used within this study is insufficient to train the neural network models to perform well with sequence tagging. In addition, multiple situations were identified in data annotation process where spelling mistakes were identified and some cases where multiple words have been merged due to typing errors. Thus, the most appropriate method to deal such scenarios were to use n-gram methods and it was clearly visible in the results as well. As future studies, the dataset is expected to be expanded with more annotated data with multiple social media data other than Facebook. Further, combinations of multiple techniques like CRF with LSTM and CRF with GRU and some other ensemble combinations are to be examined as future study with the advancement of the dataset.

ACKNOWLEDGMENT

I express my sincere thanks to LK Domain Registry for the grant given for publishing this study and to the NLP center of university of Moratuwa for providing guidelines and knowledge at each step of the study.

REFERENCES

[1] Das, A. and Gambäck, B. (2014). Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. In Proceedings of the 11th International Conference on Natural Language Processing.

[2] Vyas, Y., Gella, S., Sharma, J., Bali, K. and Choudhury, M. (2014). POS tagging of English-Hindi Code-Mixed Social Media Content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974–979, Doha, Qatar, October. Association for Computational Linguistics.

[3] Barman, U., Das, A., Wagner, J. and Foster, J. (2014). Code Mixing: A Challenge for Language Identification in the Language of Social Media. In Proceedings of the First Workshop on Computational Approaches to Code Switching.

[4] Arunavha, C., Das, D. and Mazumdar, C. (2016). Unraveling the English-Bengali Code-Mixing Phenomenon. In Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 80 – 89.

[5] Bora, M. J., & Kumar, R. (2018). Automatic word-level identification of language in assamese english hindi code-mixed data. In 4th Workshop on Indian Language Data and Resources, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 7-12).

[6] King, B., & Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1110-1119).

[7] Ramanarayanan, V., Pugh, R., Qian, Y., & Suendermann-Oeft, D. (2018). Automatic Turn-Level Language Identification for Code-Switched Spanish–English Dialog. In Proc. of the IWSDS Workshop.

[8] Malmasi, S., & Dras, M. (2015, May). Automatic language identification for Persian and Dari texts. In Proceedings of PACLING (pp. 59-64).

[9] Simon Carter. 2012. Exploration and Exploitation of Multilingual Data for Statistical Machine Translation. PhD Thesis, University of Amsterdam, Informatics Institute, Amsterdam, The Netherlands, December.

[10] Dong Nguyen and A Seza Do˘gruöz. 2013. Word level language identification in online multilingual communication. In Proceedings of the 2013 EMNLP, pages 857–862, Seattle, Washington, October. ACL.

[11] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas .1960 ;20:37–46.

[12] Smith, I. & Uthayasanker, T. (2019). Sinhala-English Code-Mixed Data Analysis: A Review on Data Collection Process. In Proceedings of the 19th International Conference on Advances in ICT for Emerging Regions (ICTer2019)

[13] Dilshani, W., Yashothara, S., Uthayasanker, T., Jayasena, S. (2018) Linguistic Divergence of Sinhala and Tamil Languages in Machine Translation. In Proceedings of 2018 International Conference on Asian Language Processing (IALP) (pp. 13-18).