SINHALA – ENGLISH LANGUAGE DETECTION IN CODE-MIXED DATA

Jude Roy Ian Smith

189350R

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

April 2020

SINHALA – ENGLISH LANGUAGE DETECTION IN CODE-MIXED DATA

Jude Roy Ian Smith

189350R

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

April 2020

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another per-son except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: ……………… Date: ……………….

Name: J. R. I. Smith

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor: …………………………. Date: ………………..

Name: Dr. Uthyasanker Thayasivam

ACKNOWLEDGMENT

I would like to express my sincere appreciation to my family for the endless support and motivation given to me to make this thesis a success. Further, I express my special gratitude to Dr. Uthayasanker Thayasivam, my supervisor, for the guidance and advices given throughout the study to successfully complete the current research. Furthermore, I would like to appreciate Dilan Sachintha, Manusha Karunarathna and Bavindu Bimsara for the assistance provided with data annotations.

ABSTRACT

Text processing is a highly demanding research area in natural language processing domain in current context. The knowledge gathered using text processing is used in variety of other domains such as artificial intelligent, optical reading, chat bots and so on. On the other hand, language detection in text has also become a trending study due to the usage of multiple languages on the internet. Further, the language identification has become a difficult function in bilingual (mix of two languages) and multilingual (mix of more than two languages) data. Accordingly, this research presents a method to detect tokens written in Sinhala and English in code-mixed data. In addition to that, this is the first such study conducted on Sinhala-English code-mixed data as per the best of author's knowledge at the time of this paper is prepared. To be precise, this is the first attempt to come up with a machine learning model on Sinhala-English code-mixed data written using Latin alphabetic characters. Indeed, if the code-mixed data is having Unicode characters, the language detection is straightforward and can be achieved using a simple Python program. However, when the whole sentence is presented in Latin characters, ambiguity increases, and it is not straightforward to detect the language and this study is a fine attempt to come up with a proper model to address this ambiguity.

In practice, Sri Lankans use Sinhala words together with English in social media platforms for communication, review posting, commenting and so on. Further, there are many methods to detect Singlish words especially Unicode characters, yet the accuracy in these models in determining Sinhala tokens or English tokens in text data (code-mixed data) are questionable. Therefore, this study presents a language detection model using machine learning and natural language processing techniques. Accordingly, two models will be introduced to identify Sinhala-English code-mixed data gathered from social media platforms and another model to identify languages in word level using the state-of-the-art techniques. In addition, the dataset of Sinhala-English code-mixed data was published in

ICTER 2019 [50] to be used for any similar studies and the final study was published in IALP 2019 held in China [51].

CONTENTS

List of tables

List of figures

1.0 INTRODUCTION

1.1 Language identification

Language identification (LID) can be defined as the method of identifying the base/natural language or the language features used in text or in a document. Further, humans can use sophisticated thoughts and ideas in communications with a known language. Likewise, humans can detect language features or simply the language used as text without a considerable effort given that that they are familiar with those languages. That is, even humans will fail to detect a language in a given text without prior knowledge of the language. Subsequently, table 1 shows the definition of natural language processing extracted from Wikipedia and its translations in different languages. Accordingly, most of the readers will identify at least one language used in the text without looking at the language type column and many readers may have the ability to recognize more than one language.

Table 1: Natural language processing description in different languages

Language	Text
English	Natural language processing (NLP) can be defended as a subfield of artificial intelligence and information engineering concerned with human and computers interaction with human languages.
Spanish	El procesamiento del lenguaje natural (PNL) es un subcampo de ciencias de la computación, ingeniería de la información e inteligencia artificial relacionada con las interacciones entre las computadoras y los lenguajes humanos (naturales), en particular, cómo programar computadoras para procesar y analizar grandes cantidades de datos en lenguaje natural
Chinese	自然語言處理（NLP）是計算機科學，信息工程和人工智能的子領域，涉及計算機與人類（自然）語言之間的交互，特別是如何對計算機進行編程以處理和分析大量自然語言數據

Sinhala	ස්වාභාවික භාෂා සැකසීම (NLP) යනු පරිගණක විද්යාව, තොරතුරු ඉංජිනේරුකරණය සහ කෘතිම බුද්ධිය යන ක්ෂේත්රයන් පරිගණක හා මානව (ස්වාභාවික) භාෂා අතර අන්තර් ක්රියාකාරීත්වය සම්බන්ධව සෙවීමය.

Accordingly, a language detection model attempts to model the ability of humans to detect a language of text. Further, there are multiple attempts those were satisfactory to produce general/specific models to address language detection. Furthermore, multiple algorithms and data structures have been created along with them. On the other hand, humans tend to use different kinds of encodings to represent characters until the ISO 8859 or commonly referred to as Unicode standard was introduced. Accordingly, with the introduction of Unicode standard, scholars were able to model language(s) within their encoding boundaries, to an acceptable level with common data structures and algorithms. However, as these models were dependent on encoding boundaries, they failed to consider the code-mixed data those are purely written in single language characters such as Latin characters.

1.2 Languages in social media

Subsequently, with the advancement of the internet and mobile phones, people are inclined to use social media platforms more frequently than the past. Likewise, most of the business activities such as marketing, recruitments, and surveys are carried out through social media resulting in people been persuaded to use social media at an increase rate. Similarly, popular social media platforms such as Facebook, Instagram, YouTube, and Twitter gather a significant amount of user data day by day. Nevertheless, it is noteworthy that there are millions of users whom will use these sites across the world using multiple languages. Therefore, a significant number of NLP researches have evolved on social media context data due to their rich nature in multiple language features and the knowledge content.

Further social media can be categorized into three categories as shown in table 2 [9]. Accordingly, it is noticeable that the type of social media varies from service to service.

Thus, the mode of communication depends on the purpose of their interactions. For instance, the mode of communication used whilst playing a game differ from the mode of communication used during a Facebook conversation. Moreover, it further differs when it relates to professional social media sites such as LinkedIn. In fact, people use Facebook for social interactions such as to post messages, share thoughts and their subsequent friends will gain the accessibility to refer them. Hence, the Facebook communications may occur in a public manner or in a private manner. In fact, private chats are used for more confidential conversations. On the other hand, social media such as Twitter is mostly used for quick communications and people tend to use short form of the words, hashtags, emoticons' and so on.

Table 2: Social media categories

Category	Example
Communications	Social Nets and Blogs
Collaboration	Social News and Wikies
Multimedia	Live Streaming, Virtual Worlds and Videos

Accordingly, these social media platforms offer a variety of services and these services were able to attract a significant amount of people towards them. Likewise, people from different geographical locations with different cultures will communicate on these platforms [6]. Therefore, due to the substantial number of users representing different languages and cultures, social media has become a new language source where different languages are created based on different purposes. These features such as acronyms are derived due to the real-time communication. For instance, LOL is used to express expressions such as laugh out loud, rolling on floor with ROFL, YOLO to represent you only live once, never mind with NVM and so on. In particular, these acronyms are prominent within real-time communication channels where people attempt to express multiple words with a single word. Further, this becomes more stimulating when people derive with different acronyms for the same expression in different languages. For an example, the acronym LOL Russians resembles using XAXA whilst Spanish people use JAJAJA [10]. In addition to that, people use symbols during communication in social media to gain attraction of others. For instance, use of hashtags (#) in posts are common as a means of gaining public attention. Yet, this character-based communication is not presented in any language vocabulary. However, people have become familiar with these characters and utilize them in many occurrences within social media postings. [23]

English language has been acknowledged as the common language used within web communications. However, there is a significant number of users whom use their native language other than English within web communications. On the other hand, most of the users whom use their native languages may fail to use Unicode characters to represent their languages. Instead, they follow phonetic typing, where they use Latin characters, and this has become popular among social media users due to simplicity. In addition to that, users add native English tokens with the mix of their native language tokens resulting in a phenomenon referred to as code-mixing (code-switching in some studies). Indeed, code-mixing is commonly recognized within peer to peer chats, group chats and group interactions such as in comments for a post. Accordingly, people may create multiple code-mixing types when they introduce Unicode characters to the communication and table 3 illustrates some of such scenarios.

Table 3 Language variation types found in social media

Variation type	Example
Purely Expressed using Latin (English) characters	We will get more algorithms to study
Purely written in Unicode	අද මම විශ්වවිද්යාලයට යනවා
Expressed using Latin characters, but with a different language (Sinhala)	Ada mama vishvavidyalayata yanawa
English characters with Unicode	අද chemistry and physics class යනවා

Code mixing	Api yaluwo kattiya film ekak balanna yanna inne

1.3 Derived languages found in social media

As explained in section 1.2, there is a natural tendency among the users to use characters from English alphabet and express their messages with these characters with their own language variations. Accordingly, this scenario has led to generate new language variations. Likewise, social media users tend to mix their native languages and Latin alphabet characters in communications, predominantly this is evitable within written communications in social media. Accordingly, new languages introduced by social media users are referred as Singlish for Sinhala language expressed in English whilst Hinglish for language which express Hindi in English and Chinglish for Chinese written using English. And these derived languages have made communication convenient in many circumstances. Further, many marketing and branding campaigns launched via SMS, email or social media also use these derived languages to reach the target segment effectively.

Currently, most of the operating systems, web browsers and smart phones support Unicode characters, hence, people can express themselves in their native language using Unicode characters. However, to use Unicode characters in a conversation, people need to pursue a general understanding of the Unicode character mapping with their device's keyboard. In addition, people may use on screen keyboards for the convenience. Yet, majority of the devices consist a keyboard with English characters. Therefore, people are persuaded to express their languages using English characters rather than using Unicode characters as stated within section 1.2. On the other hand, the users need to have an understanding on English language features and grammar to clearly understand the text written using pure English characters. In spite of the limited knowledge pursued in English grammar, the text written in derived languages such as Singlish or Hinglish may increase the understandability amongst people due to their features. Accordingly, to communicate using derived languages, the ability to read English letters is deemed sufficient. Hence, a lot of people whose native language is not English use these derived languages in their day to day communications within social media.

1.4 Language detection and its use cases

Automatic language identification (ALID) is a highly demanding research domain which aims to identify the base language(s) or the features of language(s) in code-mixed data. Further, texts are more commonly used in such studies than voice data because code-mixing is more prominent in text than in voice. However, when the data consists multiple language features, more sophisticated models are required to process such data. Yet, most of the available models fail on bilingual code-mixed data. Thus, language detection is still an open research domain where many paths are yet to be revealed.

As identified previously, social media data consists of different language features, derived languages and mixture of them and these data has made language detection trivial. On the other hand, the language detection can be easily performed if the entire text is written in Unicode characters. That is because, the Unicode standards have clearly defined the boundaries for each language segment and therefore a single decoding algorithm can easily figure out the languages and their boundaries in a given text. However, it is noticeable that the most of the social media content data are not purely written using Unicode characters.

1.4.1 Business related usage

On account of the current competitive business world, most of the companies are launching their marketing campaigns and promotions through social media. Indeed, social media covers a significant user base and it is considered as a cost-effective method of communication in comparison to other mediums of communication. Further, majority of the companies are currently maintaining their own company social media pages to interact with social media users. In addition, social media is recognized to be one of the easiest and most effective channels to collect user data and conduct surveys. Furthermore, most companies use social media to complete user feedback loop on their product or services. In fact, people review their experiences of using a product or service on relevant pages and the companies use those data in opinion mining, sentiment analysis and product survey generation. Accordingly, language detection is identified pivotal for the automation of such activities.

1.4.2 Non-business-related use cases

Language detection is not only important for business purposes, it can also be used for any regularity purposes, security purposes or even for ethical purposes. For an example, to identify whether a given text contain any hate word, first, it is important to tag each word based on their language and then carryout the evaluation. Similarly, the language detection is significantly used in various researches conducted on human behavior identification based on their communications. On the other hand, machine translations, human computer interactions are also based on language identification models.

1.5 Scope of the study

Language detection is increasing in its difficulty on a day by day basis due to the introduction of new language features by social media users. Further, it becomes more challenging when users mix multiple languages in a single communication. Thus, this study will focus only on the data generated by social media platforms and Facebook is selected to be assessed within the current research. Furthermore, the study will focus only on code-mixed data expressed using Latin characters to express Sinhala and English.

To begin with, as of present no Sinhala-English dataset is publicly available to be used to train a machine learning model. Therefore, this study will commence by creating a new dataset from the data scraped using Facebook. Further, the dataset will be annotated with multiple annotators and all relevant inter-annotator statistics will be calculated to verify the annotation. In addition, newly created dataset can be used for similar research purposes based on a non-disclosure agreement between the user and the author(s). Secondly, set of supervised machine learning algorithms will be trained with the new dataset and they will be benchmarked with the industry standard techniques for model evaluation.

2.0 LITERATURE REVIEW

2.1 Multi language learning

People are deemed to develop their own languages and language variations as they attempt to convey the target norm [37]. Accordingly, [37] interpret inter-language as a “separate linguistic system based on the observance output both errors and non-error-which result from learners attempted production of the target norm”. In addition, [13] defines some key influential factors which drive people to acquire a second language and the types of syntactic strings of the language as follows.

1. The first language background of the learner (L1).
2. The quality of teacher competence in English (L2).
3. The sociolinguistics background of the L2 learners and inherent motivation.
4. Resource availability for teaching and learning.
5. The structure of the target language (TL).

Similarly, the term inter-language is defined as the insist grammar constructed by the second language learner in the path that they learn the second language by themselves [13]. Furthermore, [37] has given five cognitive psychological processes those are central to language learning and the process exists in the latent psychological structure and that inter-language utterances are associated with one or more of these processes. Accordingly, language transfer, transfer of training, language learning strategy, second language communication strategy and overgeneralization of target language rules are the five cognitive processes given by [37].

2.1.1 Language transfer

Language transfer is the occurrence of fossilized linguistic item and grammatical rules in the language of L2 learner as a result of the background of the first language [37]. In particular, the knowledge of the first language plays a key role in acquisition of word order, interrogative and other grammar aspects, related clauses and so on. On the other hand, the deviation made by the second language learners are developmental and the transfer plays a minimal role in learning.

2.1.2 Transfer of training

Knowledge transferring is suggested as one of the most effective approaches to learn something foreign to people. Hence, knowledge transferring is significant in second language learning as well. Accordingly, the teaching variations, course content and design and mainly the teaching techniques will directly influence the learning rate and the amount of knowledge gained by the second language learner on the foreign language. Likewise, errors in each aspect can impact on the learning process in a negative manner which will result in learner to misunderstand or misrepresent the learning language [42].

2.1.3 Language learning strategy

Subsequently, language learning strategy is recognized as a similar process to knowledge transferring process with some deviations. That is, the teaching strategy or the learning strategy is imperative to acquire the foreign language by a given person. Further, it may depend on the learning capability of the person as well [47]. According to [47], simplification is the best strategy to adopt a new language and he further states that the simplification strategy will also be a reduction and modification of morphology and syntax. Furthermore, to make the language transferring convenient, this strategy includes omission of function words and plural markers.

2.1.4 Second language communication strategies

Communication is an important factor to practice and sharpen a language learned by a person [17]. Accordingly, [17] defines language communication strategies as a problem in reaching communicative milestones. Furthermore, second language learners try to adopt distinct communication strategies to facilitate the limited practice and knowledge of the foreign language [44]. Therefore, this strategy includes topic avoidance and message abandonment as a result of inadequate mastery of the transferred language. On the other hand, the learner has the capability to resort available words in the first language in combination to accomplish a communicative goal which leads to code-mixing.

2.1.5 Overgeneralization of target language rules

In general, languages have their own rules which define the language features and its usage [44]. Further, [44] states that some elements of the intended language of a learner resembles rules of a transferred language due to erroneous learning. This leads to overgeneralization and the learners tend to apply newly learned rules in an inappropriate way in sentence construction.

On the other hand, it is required to have some sort of meaningful interaction in the intended language to acquire the knowledge properly. In addition, it involves natural communication in which the learner or the communicator is concerned. Indeed, this is not with the form of the learner's utterances but the understanding and the intended message. However,[7] argues that the explicit teaching and error corrections are not relevant in language acquisition. Further, [8] adds that the language acquirers are not aware of the rules of the foreign language that they are trying to learn and may correct errors by themselves on the basis of a need for grammatical accuracy. Likewise, learning a new language is a thoughtful effort which is supported by error explicit rules and error corrections [38]. Similarly, scholars assert that the error correction is the key to language learners to overcome the challenges that they face in the path to understand linguistic generalization. Further, the good understanding of the second language and conscious learning are key to alter the final output of the second language. Therefore, this has significant importance even before the utterances are produced. That is, the accuracy of each of these points will eventually improve the accuracy of the foreign language learned [8] [38].

Intra-lingual cues from the first language and loan words form the second language generates questions such as what clues a context delivers and the effectiveness of them towards the learners learning. According to [24], children begin their language development with the patterns transmitted to them by the parents or by the care takers and any further changes are staked to that pattern. On the other hand, societal class norms influence the socialization and it is noticeable that children acquire variations favored informal speech particularly amongst lower social class. On the contrary, [31] mentions that one of the common characteristics of language acquisition is the progression from a repetitive pattern to create a language in use. Likewise, one of the uses of teaching language lexical phrases is to enable learners to acquire one or few basic or simply fixed routines. In fact, this will enable them to analyze and practice a significant number of variable patterns as they get themselves exposed to varied phrases.

2.2 Code-mixing

During past few years it is evident that there is a growing trend in bilingualism and multilingualism studies. In fact, bilingual and multilingual behavior is common within countries which practices more than one customary language. Accordingly, previous studies reveal that English language is commonly practiced in such bilingual and multilingual communications. Indeed, people practice English as a global language or as a foreign language and eventually they become bilingual or multilingual with the fluency in English. On the other hand, the countries governed by British (enclave and exclave countries) reveals bilingual and multilingual behavior than other countries due to the historical impact on their native language. Further, the advancement of internet and other devices have positively contributed towards the growth of bilingualism and multilingualism. In particular, the growth of social media platforms has a significant contribution towards this. Indeed, the social media platforms have given the opportunity for increased interaction amongst different people with diverse cultures in various geographical locations. In doing so it has necessitated people to pursue a common communication method to convey their expressions leading to generate new language variations.

The term code-mixing and code-switching has been used interchangeably amongst different studies to define bilingual behavior of humans. Nonetheless, both terms reflect a similar meaning in most of the studies to define utterances taken from multiple languages. However, in some studies, these words convey a different meaning. For instance, code-switching is referred to as the movement among languages while code-mixing is referred to as the use of multiple languages [19].

Likewise, studies have defined code-mixing as the interchange usage of different kinds of linguistic features such as affixes, words, phrases, and clauses from two or more different languages within the same sentence/speech context. On the other hand, code-switching refers to placing units such as words, phrases and sentences from two codes with the same sentence/speech context [27]. Accordingly, the key structural difference between code-switching and code-mixing is the position of the altered element in the considered context. Subsequently, in code-switching, the inter sentential alteration of the codes take place whilst the modification is intra sentential for code-mixing [27] [19]. In addition, code-mixing is deemed as a hybrid approach. For instance, distinct grammar elements from different grammatical systems are used within code-mixing whilst code-switching emphasizes the movement from one language to another by a bilingual/multilingual speaker [27]. In other words, code-switching emphasizes linguistic performance whereas code-mixing emphasizes the formal aspects of linguistic competences.

Similarly, [18] defines code-mixing as the practice of using multiple language features in a single communication attempt. Further, this phenomenon is more highlighted in text and voice communications. Furthermore, code-mixing practices are more common in social media communication as well as in online review posting [4]. In fact, code-mixing is evitable when people are familiar with more than one language and they tend to use one or more foreign language with their base language. However, in practice, the maximum number of languages those are mixed by a person will be three, yet, it is common to observe mixing of two languages. That is, most of the time, the code-mixing limits to one foreign language with the base language [18]. Accordingly, the usage of two languages is identified as bilingual whereas trilingual is the term used for usage of three languages. Similarly, usage of more than three languages will fall into multilingual category. Further, most of the studies are conducted considering only the bilingual scenario due to its dominance in communication [18].

2.2.1 Motivations of code-mixing

As stated earlier, code-mixing is a common phenomenon which can be seen among people whom are familiar with more than one language. Accordingly, scholars have identified some motivational factors which has driven people to practice code-mixing within their communications.

1. **Sociolinguistic approach**

Sociolinguistic factors as a motivator for code-mixing have been studied by [3] in a small Norway community. Accordingly, the study segmented code-switching into two main categories namely situational switching and metaphorical switching. The main cause for situational code-mixing is the differences of users' social settings. Further, metaphorical code-switching arises due to the topic change which ultimately leads to change in speaker's identity. On the other hand, [19] defines the term conservational code-mixing and records six probable aspects which leads it to take place: addressee specification, quotation, reiteration, interjection, personalization, and message qualification. Further, within this study code switching performed in a single language is referred to as 'they-code' whilst the term 'we-code' is used to define a sense of solidarity and personal involvement. Moreover, the term 'they-code' refers to represent a detachment from a certain group.

In addition, [14] has conducted a study on community languages and presented eight factors which lead to code-mixing practices: role relationship, interlocutor, domain, venue, topic, channel of communication, phatic function, and type of interaction. Further, it states that there is a high influence by the speaker's motivation in defining the use of each language. For instance, the language usage of the speaker will depend on the lifestyle and the fluency of the language.

In addition, in a study conducted by [29] has recognized a model named marked model which records that the code-mixing can be used to exhibit different identities of the speaker. Accordingly, another study has been conducted in Kenya and other African countries, where the scholar suggested that the choice of a language can be marked or unmarked in practice. In marked code selections, the speaker intends to convey set of unpredicted privileges and responsibilities between the speaker and the addressee.

2. Linguistic approach

In addition to sociolinguistic motivators, linguistic approaches have also considered as an important factor which drives code-mixing. Accordingly, [36] defines two language constraints in code-mixing as equivalence constraint and free morpheme. In fact, equivalence constraint states that code-mixing is not possible if the user fails to follow syntactic rules of both the languages.

Further, code-mixing can be sub divided based on linguistics and according to [34], there are three different kinds of code-mixing namely: intra-sentential mixing, inter-sentential mixing, and tag mixing. Furthermore, each of these topologies of code-mixing indicates the competences of the speaker [34]. On the other hand, [28] proposed an influential model called matrix language model and the two languages involved in a code-mixed communication were categorized as an embedded language or matrix language. Thus, the matrix language will determine the morph syntactic frame of the code-mixed route and it will also provide system morphemes. On the other hand, embedded language provides the content morphemes. Further, number of scholars have studied about descriptive analysis of code-mixing languages. [11] has studied code-mixing on Hong Kong’s context, Cantonese English, and the majority has identified English nouns, followed by verbs and adjectives as the most common code-mixed syntactic category. However, the findings of [11] is inconsistent with some other studies. For instance, [40] has studied Kannada-English code-mixed data and found that adverbs and adjectives were frequently code-mixed than verbs. Similarly, [20] identified based on a Moroccan Arabic code-mixed data that the most frequently code-mixed elements were nouns and discourse makers.

3. Code-mixing in different genres

As reported by many scholars, different genres also influence the code-mixing practices other than sociolinguistic and linguistic factors. Accordingly, [34] suggests that the genre analysis done by [41] was useful on a study done in Hong Kong to demonstrate the likely viewpoints for bilingual conditions. Further, genre analysis allows to identify smaller domain of languages used, to define languages used by specific groups, and potentially to mix language in each of the similar domains [32]. Further, [32] suggested that the genre analysis is useful in professions and in domains where Cantonese-English code-mixing is used. On the other hand, [12] has analyzed the features of code-mixing in Hong Kong Cantopop and used theories related to genre and decided that driving factors to account for Cantonese-English code-mixing lacked proper translations. In particular, they were insufficient to explain the reasons of having English elements within the lyrics of some of their songs.

Further, [11] argues that the code-mixing is less frequent in text communication than in oral communications. This is because, within text communications, the users get the time to think and generate a message in any language and translate the message into another language. However, in oral communications it is a spontaneous process which will create real-time pressure to express the ideas quickly. Likewise, this study further identified that there were written communication and oral communication code-mixing which may carry different functions. For instance, [26] discusses that code-mixing in texts are more of a stylish element whereas code-mixing in voice communications serve wide range of functions. However, he also argues that the difference between both code-mixing is not clear in some practical scenarios. According to [43], literacy: properties of written conversation and orality: properties of vocal dialog, are two tails of the same continuum. Therefore, functions from these two practices may exchange. That is, there are some voice communicational characteristics found in written communications and some written communicational features were found in voice communication and vice versa [11].

2.2.2 Myers-Scotton’s model

As introduced by Myers-Scotton, code-mixing practice named as markedness model has been cited significantly on code-mixing studies [30]. Accordingly, this model states that the code-mixing is a strategy which is motivated by social aspects employed for producing a sequence of unmarked choices, to establish itself as a marked choice. Further, speakers switch between languages when they initiate a conversation in an unmarked choice. On the other hand, the name markedness relates to the choice of one linguistic variety from other possible varieties. Furthermore, the model also classifies code-mixing into four different categories as: marked, unmarked, sequential, and exploratory code-mixing [30].

1. Code-mixing as marked choice

Marked code-mixing leads speakers to make a choice in codes [29]. Rights and obligations (RO) are a theoretical aspect which helps speakers to base their expectations in each setting [29]. Further, RO is responsible for codes of norms and behavior those are accustomed and practiced within societal communities. Likewise, such a selection will be made by the speaker when he or she intends to create a new RO set which is unmarked for the current exchange. Further, in code-mixing as a marked choice, speaker demands to maintain the distance between his or herself and with the expected RO [29]. In fact, the marked choice will occur in formal conversations for which an unmarked language choice is expected by the participants. Likewise, the choices made by a speaker clearly indicate the appropriate RO set in their social context. On the other hand, it can also be stated that marked choice is a negotiation against the unmarked RO. Furthermore, one reasoning for the speakers to engage in a marked code-mixing is to indicate a diversity of emotions from affection, anger, surprise, sadness and so on. Further, marked code mixing is deemed as a conciliating factor of the authority demonstration as well as a declaration of the ethnic identity.

2. Code-mixing as unmarked choice

The unmarked code-mixing choices drives speaker to involve in communication in a certain way based on the situation. Further, unmarked code-mixing occurs mostly in multilingual communities when two or more languages are spoken in a single conversation [30]. On the other hand, conditions which drive unmarked code-mixing differ from one community to another. Similarly, in African communities, code-mixing is done between colonial languages and indigenous languages. Further, Africans use their own customary language with their peers from the same tribe and English or Afrikaans with the other African communities. In addition to African communities, this practice can be seen in South Asian region especially in the case of people originating from rural areas whom are only familiar with their native language only or/and English. Even political and economic aspects drive code-mixing in these communities and the matrix or the base language of the local conversations is normally conducted using their native language than using English or Afrikaans [29].

3. **Code-mixing as an exploratory choice**

Subsequently, the speaker may use exploratory choice when they are unclear on the unmarked code choice. Accordingly, when the speaker is uncertain on the expectation of RO or the optimal communicating intent it is deemed that this type of code mixing may take place. Further, exploratory code-mixing is uncommon in practice and it is not often used as an unmarked choice [29].

2.2.3 Code-mixing types found in social media

In practice, different kinds of code-mixing can be identified within social media, namely, data as intra-sentential code-mixing, inter-sentential code-mixing and tag switching.

1. **Intra-sentential code-mixing**

Intra-sentential code-mixing is one of the code mixing types found in social media data. Indeed, this code-mixing type was introduced by [34] and the code-mixing happens at the level of words within sentences. Further, the code-mixing may occur in the middle of a clause, sentence or even between words. According to [21] intra-sentential is the most frequent type of code-switching found in bilingual communities.

2. **Inter-sentential code-mixing**

On the other hand, inter-sentential code-mixing occurs only within a clause boundary. In other words, this type of mixing is visible at the clause, phrase level or at word level if there are not any morphological adaption occurrences. In addition, inter-sentential mixing is also suggested by [34] and it was defined that “inter-sentential switching involves a switch at a clause or sentence level in different languages”. Accordingly, switching either within the clauses or between sentences in one language may conform to the rules of the other language.

3. **Tag switching**

The last type of code-mixing suggested by [34] is the tag switching. Accordingly, this involves the switch of a tag phrase or a word or both from one language into another language. Further, these tags may be placed in a sentence as desired by the user or in an utterance which are in the other language.

2.3 Related work

Subsequently, with the high usage of social media platforms, a new area of natural language processing was emerged to process social media data. Accordingly, due to the multiple language usage in social media platforms, language identification in code-mixed data has become one of the tedious tasks. Moreover, this is acknowledged as one of the highest demanding area of study. Even though the language identification (LID) in text data is significantly researched, this is not evidenced in studies related to code-mixing. However, multiple code-mixed data classification attempts can be found mainly on languages used by Indians and Bengalis with English. [15][46] have studied code-mixing in Bangla and English while [2] studied on Hindi and English code-mixed data. Despite of the high demand for this research area, the empirical evidence suggests that the word level language detection as a difficult and a noisy process and may found in most of the social media related work.

Accordingly, [46] has made a significant contribution to identify the language detection in word level on code-mixed data generated by social media platforms. Subsequently, the main languages studied in this research are Hindi-English code-mixed data and Bengali-English code-mixed data. Likewise, the study has used multiple machine learning methodologies to address the language identification task and a dictionary-based classifier has used as the baseline model. Likewise a SVM classifier with four features namely; dictionary features (binary feature for each language generated based on the presence/absence of the each token in dictionary of that language), weighted char-n-grams

(3-grams and 4-grams), minimum edit distance weight (for out-of-dictionary words) and context information for each token (3 previous token labels and 3 following token labels) were used for the study. Accordingly, the study has got a high precision value higher than 90% on Hindi-English code-mixed data and 87% precision for Bangla-English code-mixed data. Yet, the model gives a low recall value of 65% and 60% for Hindi-English and Bangla-English, respectively. This is because; the model does not perform well predicting actual labels.

Further, [15] presents a study conducted on Hindi-English code-mixed data on social media and it also introduced a POS (part-of-speech) annotation model. As a result, matrix language of the whole sentence and the fragment language of each token has been annotated with the POS model and the study used 3,218 Hindi token and 3,210 English tokens extracted form SMSs. Hence, this study produced a logistic regression model resulting a F1 score of 0.87. Nonetheless, the model has performed poorly when considering the recall value for Hindi tokens.

As given in [1], a code-mixed data analysis has been carried out on Bangla-English data using Facebook chat history data and with another dataset from FIRE 2013. Accordingly, as features of the best performing mode, they have used dictionaries for both languages, n-grams, surrounding words to train the model. As a result, the model has been able to achieve a F1 score of 91.5% for both datasets and it outperforms the model introduced in [1]. However, the used methodology and the experiment setup is vague and failed to use the state-of-the-art methodology.

In addition, a study done by [2], has used a refined methodology to detect languages in Indian language code-mixed data. Thus, when compared to other studies, this research has studied multilingualism with a dataset of Hindi, Bangla, and English languages. Hence, the dataset contains 9,813 Facebook comments and 2,335 Facebook posts and has been annotated with sentence, inclusion, fragment and wlcm (word level code-mixing) tags. Further, the study produced a CRF model and an SVM model trained with 3 dictionaries for each language, char-n-grams, word length, context information, and capitalization features. As per the results, CRF model has an accuracy of 95.76% while SVM shows 95.52% accuracy.

Similarly, [5] has tested decision tree model and an SVM model to be used with language detection on Assamese, Hindi, and English code-mixed data. This study has also selected Facebook as the data source and collected 4,768 tokens resulting in 20,781 tokens. The study has used word unigrams and prefixes/suffixes, word unigrams and context information as features to train each model. Accordingly, the decision tree model has got the highest accuracy of 96.01% thereby outperforming the SVM model.

On the other hand, [22] has conducted a language detection study on a dataset which consists of 30 languages taken from publicly available data from web pages. Further, the study has tested three models with six different features. Likewise, a logistic regression model trained with generalized expectation (GE), HMM model trained with expectation maximization (EM) and finally a CRF model trained with GE were introduced. As features, full word, char unigram, bigram, trigrams, 4-grams, and finally 5-grams have been used to train each of the model. Accordingly, the results reflected that, CRF model has scored the highest accuracy compared to other models, yet, less than 95% which is less than the model introduced by [5].

In contrast, [35] has used a different approach in data collection for the language detection study and it has used voice data of English and Spanish and translated them into text data resulting in 242,475 tokens. Similar to most of the other studies, this also has used char n-grams, word n-grams and character prefixes/suffixes as features and trained a CRF mode, a LSTM model, and a logistic regression model. Finally, the CRF model has outperformed other two models which were trained with char n-gram and word n-gram features with 91.0% accuracy.

In [25] code-mixed dataset of Persian and Dari languages have been studied with 28,000 Dari sentences scraped from American news websites. In this study, only word n-gram and char n-gram have been used as features to build an SVM classifier. The best performing SVM model accounts 96% as the accuracy score and the authors have used another Uppsala

Persian Corpus which was completed out of the domain dataset to evaluate the model's generalizability. Finally, the model has classified 79,000 sentences with an accuracy of 87% even with out of domain corpus.

Unlike in other studies, [39] has selected Twitter social media platform to study code-mixed of five European languages resulting in a considerably larger dataset of 1.1 million tweets. Further, weighted n-gram has been used to train the model and it revealed an accuracy of 92.4%. Similarly, [16] has produced a model for word level language detection on a dataset of Turkish-Dutch speakers. Further, a single annotator was used to annotate the corpus and only a randomly selected 100 posts were annotated by a second annotator to calculate inter-annotator statistics. As the study reports, the Kappa statistics for inter-annotator agreement is 0.98 which is a well agreed annotation task between annotators. Accordingly, 3,066 posts, 29,385 Dutch tokens and 26,170 Turkish tokens created the whole training corpus. Subsequently, the study has tested a logistic regression model, a dictionary-based model, and a linear chain CRF model. Finally, a high accuracy of 97% has been achieved on the token level with a lower accuracy of 89.5% in sentence level.

In contrast, [45] has studied Romanized Arabic dialect in code-mixed tweets to identify languages in token level. The study has been done with two main models named LDA (Latent Direchlet Allocation) and ME (Maximum Likelihood) model. Accordingly, the corpus used for study was created using tweet data and 475,338 tweets has been used for LDA model and 80,131 tweets to train the ME model. All the tweets have been annotated manually by a single annotator based on the language. Indeed, the LDA model is a state-of-the-art unsupervised learning model which is heavily used for text classifications. Furthermore, the authors have used several LDA models for different topics of data and the BOG (Bag Of Words) techniques have been used for all the model trainings. The study also consisted of a supervised machine learning model using DATool as well. Finally, the best performing model of this study has recorded a precision of 95.5% for English, 95% for French and 73% for Darija. Further, the final model results a recall of 93.25% for English, 90.1% for French and 83.4% for Darija. Accordingly, the results of the study lack in precision as well as in recall for Darija even though the model perform well for the other two languages. However, the main drawback of the study is deemed as the class distribution of each language tokens where English tokens are nearly half of the French tokens.

Table 4: Recent studies in automatic language detection since 2012

Paper title	Published year	Studied Languages	Dataset(s)	Models(s) used	Features used	Results
POS tagging of English-Hindi Code-Mixed Social Media Content	2014	• Hindi • English • Bengali	6,983 posts and 113,578 words from Amitabh Bachchan, Shahrukh Khan, Narendra Modi, and the BBC Hindi news page	• Dictionary based classifier – baseline model • SVM	• Word context information • Weighted character N-grams • Minimum edit distance weight • Dictionary feature	• 90% precision and 65% recall for Hindi-English • 87% precision and 60% recall for Bangla-English
Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text	2014	• Hindi • English	3201 English words scraped from SMS data and a Hindi corpus of 3218 words	• Word-level logistic regression model	• N-gram with weights • Dictionary feature • Minimum • Word context information edit distance	• F1 score of 87%

Unraveling the English-Bengali Code-Mixing Phenomenon	2016	• Bangla • English	FIRE 2013 dataset- 539 tokens and 518 Facebook chat data	• Decision tree • K-nearest neighbor	• N-gram with weights • Dictionary feature • Surrounding word label	• F1 score of 91.5% for Facebook chat dataset and 91.5% for FIRE data set
Code Mixing: A Challenge for Language Identification in the Language of Social Media	2014	• Hindi • English • Bengali	9,813 Facebook comments and 2,335 Facebook posts	• CRF • AVM classifier	• N-gram with weights • Dictionary feature • Word length • Capitalization • Contextual information	• Accuracy of 95.76% for CRF • Accuracy of 95.52% for AVM
Automatic word-level identification of language in Assamese English Hindi code-mixed data	2018	• Assamese • English • Hindi	4,768 Facebook comments	• Decision tree • SVM	• Word unigrams, word unigrams • Prefixed and Suffixes • Contextual information	• 96.01% accuracy for SVM

Labeling the languages of words in mixed-language documents using weakly supervised methods	2013	• 30 languages	252,360 lines	• CRF with GM • HMM with EM • Logistic regression with NB	• Character unigrams • Bigrams, trigrams • 4-grams • 5-grams	• Accuracy is less than 95%
Automatic Turn-Level Language Identification for Code-Switched Spanish–English Dialog	2018	• Spanish • English	242,475 words of text	• CRF • Logistic regression • LSTM	• Word n-grams • Character n-grams • Character prefixes and suffixes	• CRF model with an accuracy of 91%
Automatic language identification for Persian and Dari texts	2015	• Persian • Dari	28,000 Dari sentences from an American news website	• SVM	• Word n-grams • Character n-grams	• 96% accuracy
Exploration and Exploitation of	2012	• Five different European languages	1.1 million tweets	• CRF	• Weighted n-grams	• 92.4% accuracy

Multilingual Data for Statistical Machine Translation						
Word level language identification in online multilingual communication	2013	• Dutch • Turkish	29385 Dutch tokens and 26170 Turkish tokens	• Dictionary based mode • Logistic regression • Linear-chain CRF	• Dictionary feature • Character n-gram •	• 89.5% accuracy on post level
Finding Romanized Arabic Dialect in Code-Mixed Tweets	2014	• English • Arabic • Romanian (Latin)	475,338 tweets	• LDA	• Bag of words	• English (P: 95.5%, R: 93.25%) • French (P: 73%, R: 83.4%) • Darija (P: 73%, R: 83.4%)

3. METHODOLOGY

As stated earlier, the current study intends to create a method with machine learning methodologies to be used for the identification of languages in each text segment. In spite of the multiple code-mixing analysis studies conducted by scholars on various language mixes, there are not any published studies found on automatic language detection in Sinhala-English code-mixed data based on the author’s knowledge by the time of this study is conducted. Hence, as per the author’s awareness, the current study is identified as the first attempt to evaluate the Sinhala-English code-mixed data and automatic language detection within Sri Lanka. To be precise, this is the first attempt to come up with a machine learning model on Sinhala-English code-mixed data written using English alphabetic characters. Accordingly, as this is the first study on above scenario, datasets were not available to carry out such a study. Thus, as the first step of this study, new dataset was created to be used to create a model to detect the language mix in text. Furthermore, the model development is divided into two parts to ease the complexity of the current study and hitherto a model to classify code-mixed data and another model for sequence tagging was conducted as two areas of study.

3.1 Data collection

Subsequently, to create the code-mixed dataset, social media was chosen as per the literature reviewed. Indeed, empirical evidence revealed a higher number of code-mixing activities using social media. In particular, the data collection was further narrowed down to Facebook which consists of a higher amount of code-mixing communications. Accordingly, publicly available Facebook posts, comments and a few of Facebook chat data was collected for the study using Facebook API. Likewise, the dataset collection ended up with 7,500 lines/sentences which resulted in 40,915 tokens after a manual cleaning process. Subsequently, the data consisted of Singlish Unicode tokens, Sinhala tokens, English tokens, and emoticons. Therefore, the dataset had to be pre-processed manually to filter only the sentences with Singlish tokens, English tokens, or Sinhala Unicode tokens.

3.1.1 Annotations

In particular, to use this dataset in a classification study, it is necessary to assign labels based on the context which suites for the purpose of the study. Hence, the entire dataset was annotated sentence wise as well as token wise with appropriate labels in two different annotation processes. The first annotation process is named as level 1 annotation and the second phase was identified as level 2 annotation. Accordingly, all the annotations were manually performed by three computer science and engineering undergraduates whom were native Sinhala speakers. However, they were fluent in English language as well. Furthermore, each annotator was paid based on the annotation process. As the annotation media, Google sheets were used to carry out all the annotations and each of the annotator was given separate datasets in separate Google sheets and they were asked to annotate the given data remotely. Further, for a given data batch, each annotator was given a batch of data which did not overlap with another annotator's data set. Likewise, the full dataset was assigned in a rotational manner and at each step, the dataset was shuffled before assigning to improve the annotation quality and eliminate any collaborative work between annotators.

1. Assumptions

As in any other communication, social media communication also consists errors on spellings. This is because; the communication in social media is deemed as near real time communication between peers. In addition, the social media users tend to shorten words to make the communication/typing faster which contributes to significant amount of errors related to spellings. Accordingly, table 5 illustrates the frequent short form words found in the dataset and their most probable complete word. Nevertheless, the usage of short words is deemed unavoidable and mostly these words contributed to code-mixing. Thus, the annotators were requested to neglect any short form words or spelling mistakes in each token and assign the labels based on their most probable complete word.

Table 5: Short form tokens and their probable complete token

Short form token	Most probable complete token
Tnx	Thanks

Flm	Film
Tkt	Ticket
Plz	Please
Thnk u	Thank you

2. Annotation quality evaluation

Subsequently, the dataset was evaluated to quantify the quality of the data preparation and statistical methods were used in most of the situations. Likewise, the agreement between annotators directly terminates the quality of the final data annotation. Therefore, the percentage of total agreement or the percentage of effective agreement were used to evaluate the level of agreement between annotators. However, this method does not account for the agreement occurred by chance. In particular, Cohen’s Kappa measures the trustworthiness of an agreement. Indeed, the agreements those are beyond the expected by change are deemed positive [48]. Likewise, Kappa statistics provides the percentage of agreement occurred beyond by chance. Thus, all the annotation phases were evaluated with Kappa statistic calculations to measure the annotation quality (eq (1) and eq (2)).

3. Level 1 annotation

First annotation phase was dedicated for sentence level annotation. Accordingly, depending on the language(s) used in each sentence a label was assigned. Table 6 shows all the labels used in the annotation with the examples. Accordingly, the same table was given to all the annotators before the annotation.

Table 6: Tags used for level 1 annotation

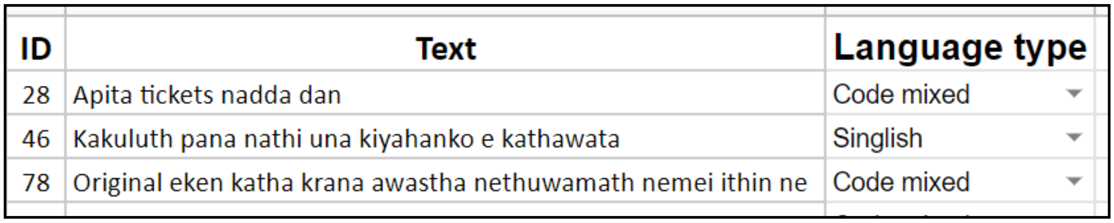
Language type	Description	Example
English	All the words are written in English language	Good morning
Singlish	All the words are written in Latin characters but represents words in Sinhala language	Mama ennam

Sinhala (Unicode)	Each word is written in Unicode characters	සුබ උදෑසනක්
Code-mixed	Unicode characters mixed with Singlish words	මම yanawa
	mama ennam film hall ekata	Sinhala words written using Latin characters (as Singlish), but with mix of some words from English language in the sentence
	Unicode characters mixed with English words	Dreams පුදුම හිතෙනවා
	Unicode characters mixed with English words and Singlish words	පුදුම හිතෙන dream ekak
Unknown	Anything which does not falls in any category stated above	

Accordingly, the full dataset with 7,500 lines was split into three batches of 1,500, 3,000 and 3,000 sentences. Afterwards, each annotator was given a Google sheet with a batch of sentences which was not overlapping with each other. Likewise, three batches were rotationally assigned to each annotator and each of the batch were annotated by all three annotators resulting in each sentence been annotated three times by all three annotators. On completing the annotation process, all the annotations were aggregated, and final annotation was selected based on the total agreement for the sentence. That is, to assign a label for a given sentence, all three annotators should have agreed on the same label and all the sentences which failed to achieve a total agreement of the annotators were discarded.

As shown in table 8, table 9 in count wise and in figure 2 in graphically, Singlish sentences dominate the dataset with 4,691 sentences and there are 1,900 code-mixed sentences which were useful for the sequence labeling study. On the other hand, the number of English sentences were 476 which is 6.34% of the total dataset and it is negligible compared to other languages. Therefore, it can be stated that, based on the collected dataset, Sri Lankans use Singlish (mostly) in their social media communication (Facebook in this study) compared to other languages. Further, there is a considerable number of sentences (25.33%)

which were identified as code-mixed and it is evident that the users have used language mixes to a considerable extent. And also, it is noteworthy that out of 7,500 sentences there were only 7,080 sentences which achieved the total agreement whereas 420 sentences failed to achieve the same label from all three annotators.



ID	Text	Language type
28	Apita tickets nadda dan	Code mixed
46	Kakuluth pana nathi una kiyahanko e kathawata	Singlish
78	Original eken katha krana awastha nethuwamath nemei ithin ne	Code mixed

Figure 1: Sample of level 1 annotation process

Table 7: Languages mix after level 1 annotation

Language type	**Batch 1**	**Batch 2**	**Batch 3**
Singlish	926	1,883	1,882
Code mixed	342	720	838
English	89	211	176
Unknown	0	5	8

After the level 1 annotation phase, annotation quality was calculated using Cohen’s Kappa statistics (eq (1)) to measure inter annotator agreement. Accordingly, the equation provided in eq (1) was adopted for this study as shown in eq (2). Likewise, Kappa values were calculated for each batch separately and for the full dataset using eq (2). And as shown in table 8, all three batches have scored a higher kappa values indicating high agreement between all three annotators in each batch. Further, the final Kappa value for the full annotation phase resulted in 0.88772595 rationalizing a higher agreement between the three annotators for the full dataset.

$$K = \frac{P_{agree} - P_{chance}}{1 - P_{chance}} \quad \text{------------------------} \text{ eq (1)}$$

Where,

P_{agree}= proportion of trials which judges agree

P_{chance} = proportion of trials in which agreement would be expected due to chance

$$K = \frac{P_{total\ agreement} - P_{chance}}{1 - P_{chance}} \text{ ------------------------ eq (2)}$$

Where,

$P_{total\ agreement}$ = proportion of sentences where all three annotators assign the same label

P_{chance} = proportion of sentences in which different labels assigned by annotators

Table 8: Kappa values (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

Table 9: Total language wise sentence count

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

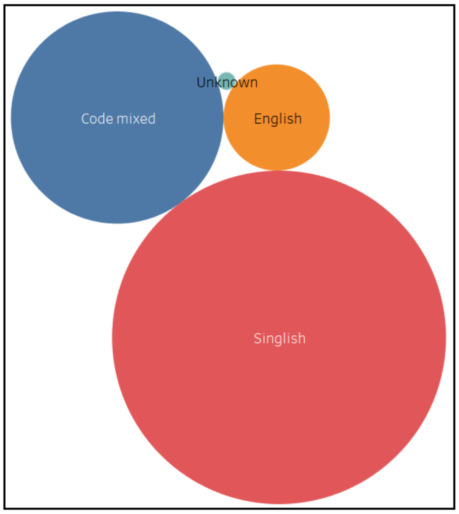
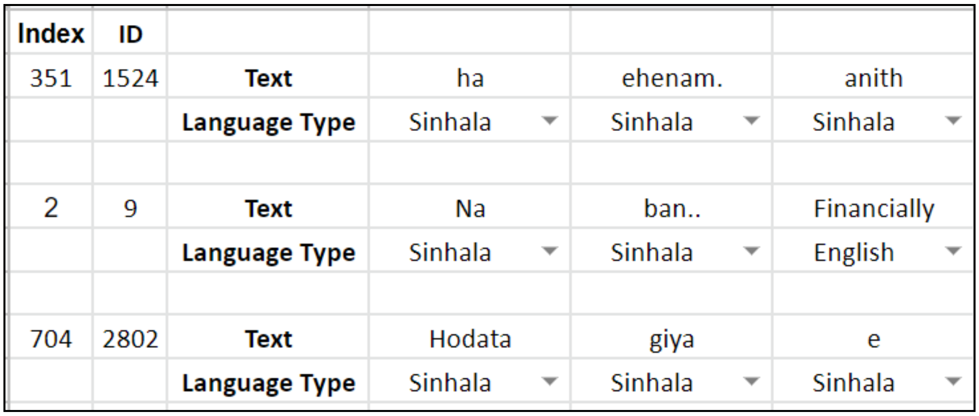


Figure 2: Graphical representation of total language wise sentence count

4. Level 2 annotation

Accordingly, within this annotation phase, each selected sentence from the first annotation process was annotated in token level. In level 1 annotation, 1,900 sentences achieved complete agreement as code-mixed and those 1,900 sentences were used for the second level annotation. This is because, the aim of the current study is to detect each language used in each text and only code-mixed data consists of more than one language were used for this phase. In similar to the first annotation, word level annotation was performed using Google forms. However, as the study is to detect Sinhala and English languages in a text, it required only two labels to be presented within the dataset. Nonetheless, within the code-mixed data, there were name entities, numbers and some special characters presented within different locations in some sentences. On the other hand, some of these tokens were ambiguous to label as Sinhala or English and therefore two additional tags were used referred as "Name" and "Unknown". Name tag was used to annotate name entities and all the other tokens which do not fall into Sinhala, English or Name category were annotated as Unknown.

However, unlike in the first annotation process, the total dataset was assigned as a single batch to each annotator. Further, dataset was shuffled before assigning to an annotator and therefore each annotator received a different sentence order on each occasion. As illustrated in fig 3, each sentence was given with separate set of slots for each token and annotators were asked to select a label in the given slot. As in the first annotation phase, annotators were asked to neglect any spellings mistakes, short form words or acronyms and annotate each token with their base language. Moreover, the label of a token was not decided only based on its base language, hence, the annotators were requested to annotate each token based on the base language of the surrounding tokens. This is because, even though a token is recognized as English or a Sinhala word at the initial outlook, the actual language may be different when evaluating the surrounding words. For an instance consider the sentence “me ahanna”. The word “me” resembles an English word for “me/myself” in the first glance, while the second token is a Sinhala word written in English characters which represent the meaning “listen” in English. Therefore, the first token “me” is a Sinhala word presented hey in English. Thus, if the first token is annotated as English and the second token as Sinhala, the complete sentence will be meaningless. This is because, the original language of the first token is Sinhala. Therefore, before assigning a label for a token, surrounding tokens were considered.



Index	ID				
351	1524	**Text**	ha	ehenam.	anith
		Language Type	Sinhala	Sinhala	Sinhala
2	9	**Text**	Na	ban..	Financially
		Language Type	Sinhala	Sinhala	English
704	2802	**Text**	Hodata	giya	e
		Language Type	Sinhala	Sinhala	Sinhala

Figure 3: Sample of Level 2 annotation sheet

Likewise, 11,795 words from 1,900 sentences were annotated with the base language whilst accounting for surrounding words. Further, unlike in the first annotation phase, the final label of a token was selected based on the majority annotation. This is because, if the final annotation was decided based on the total agreement, some tokens would have to be removed from sentences if they failed to achieve total agreement and it may discard the meaning of the full sentence. And as per the majority voting decision, there were no instances which could not be assigned to a final label. On the other hand, the Kappa statistic of this phase also resulted as 1 because there were not any random chances of selection.

As shown in table 10 and in figure 4, Sinhala tokens dominated the code-mixed dataset resulting in 8,568 tokens and users are using Sinhala tokens significantly in code-mixed communications which is an expected result when considering the first level annotations. Further, this was further expected due to the dominance of Singlish sentences within the full dataset which accounts to about 4,691 sentences out of 7,500. On the other hand, there were 2,824 English tokens which was 23.94% from the total token count and it can be concluded that on average (23.94% ~ 25%) 1 out of 4 tokens is an English token whilst the others are Singlish.

Table 10: Summery of Level 2 annotation

Annotation	**Token count**
Sinhala	8,568
English	2,824
Name	350
Unknown	53

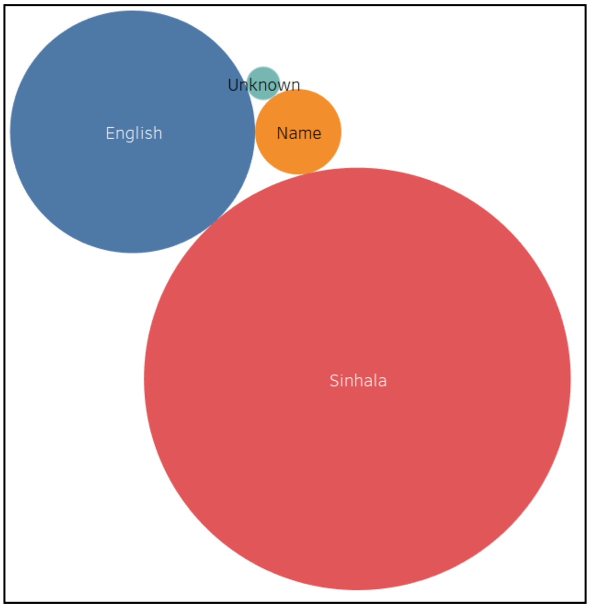


Figure 4: Graphical representation of level 2 annotation statistics

3.2 Dataset analysis

3.2.1 Frequent words

Since the study is concentrated on code-mixed data analysis, only code-mixed data selected from first annotation process were further analyzed. Accordingly, there were 3,655 unique words found in the code-mixed dataset with 2,620 unique Sinhala words and with a unique English token count of 1,035. On the other hand, when considering the token count as shown in table 12, top three frequent tokens carried a similar meaning in some scenarios. Indeed, this may be because the users have used these two words interchangeably. However, as shown in table 13, prioritized English tokens carried different meanings.

Table 11: Unique token usage

Language	Total token count	Unique token count	Percentage (unique count/total count)
Sinhala	8,568	2,620	30%
English	2,824	1,035	36.65%

Table 12: Frequent Sinhala words

Word	Word count
eka	353
ekak	253
eke	138
ne	116
ekata	85
kiala	82
hari	73
api	70
mata	66
man	66

Table 13: Frequent English words

Word	Word count
set	189
call	47
scene	39
sure	37
trip	37
car	37
film	33
van	32
group	27
sorry	24

3.2.2 Ambiguous words

Subsequently, when users express their native language using a foreign language, there is a possibility that the expressed token may resemble similarities to the foreign language.

Hence, they are denoted as ambiguous words within the current study. In most cases, ambiguous words are difficult to annotate without context clues and in some cases, they are difficult to annotate even with the context clues. Therefore, this phenomenon also led to take the majority vote to decide the final label of a token in the second annotation phase.

Ambiguous words were selected when there is no complete agreement for a token and the most frequent ambiguous tokens were found in the code-mixed dataset are shown in table 14. Afterwards at the end of the annotations, a discussion was held with the annotators to identify the root causes lead to these ambiguities. Accordingly, the main causes recognized were some words did not carry a clear definition and some sentences lacked the surrounding words. As an example, the word “Sinhala” is ambiguous because it was not clearly defined whether it is a Sinhala word itself or is it an English word for the language itself. However, in the annotation, the majority vote is for Sinhala and therefore it is labeled as a Sinhala token. On the other hand, the word “Royal” is also ambiguous because it is a name of a Sri Lankan school as well as a token from English vocabulary. Thus, if the context clues are limited, then it is difficult to determine whether it is the Royal collage or the English token Royal.

Table 14: Ambiguous words

Token	**Annotation**	**Annotator justification**
sinhala	Sinhala	Sinhala word itself is a Sinhala word used to identify the language
oke	Sinhala	Sinhala term to indicate “that” in English
royal	Name	Represent Royal collage in Sri Lanka
sup	English	A short term to represent “Support” term in English
maxaa	Sinhala	Expression of complement used in common communications

shape	English/Sinhala	The word “shape” specifies its base English meaning which is the form of an object. But, in some context in communication, it is used to express the word “nevertheless” in Sinhala.
okkkk	English	Represents the English word “Okay”
prenzz	English	Short form of the term “friend” in English
10k, 5k and 15k etc.	Sinhala	Indicates “exactly 10, exactly 5 and exactly 15” even though it looks like presenting 10,000, 5,000 and 15,000.
5i, 6i	Sinhala	Indicates “exactly 5, exactly 6” in Sinhala

3.3.3 Word2Vec representation

Finally, all the tokens were visually inspected to check whether there are clear segments of tokens in language wise or when represented all tokens together. Likewise, Word2Vec library was used from Genism package to represent each token by a vector with a length of 300 and the word2Vec model was trained on the code-mixed dataset. Figure 5, 6 and 7 illustrated these word vectors projected into two dimensions for graphical representation. Subsequently, when considering the graphs in 2D space, all most all the words forms a single segment which is highly concentrated at the center of the cluster irrespective of the language. Thus, it can be concluded that it is required to have more dimensions to identify each word individually.

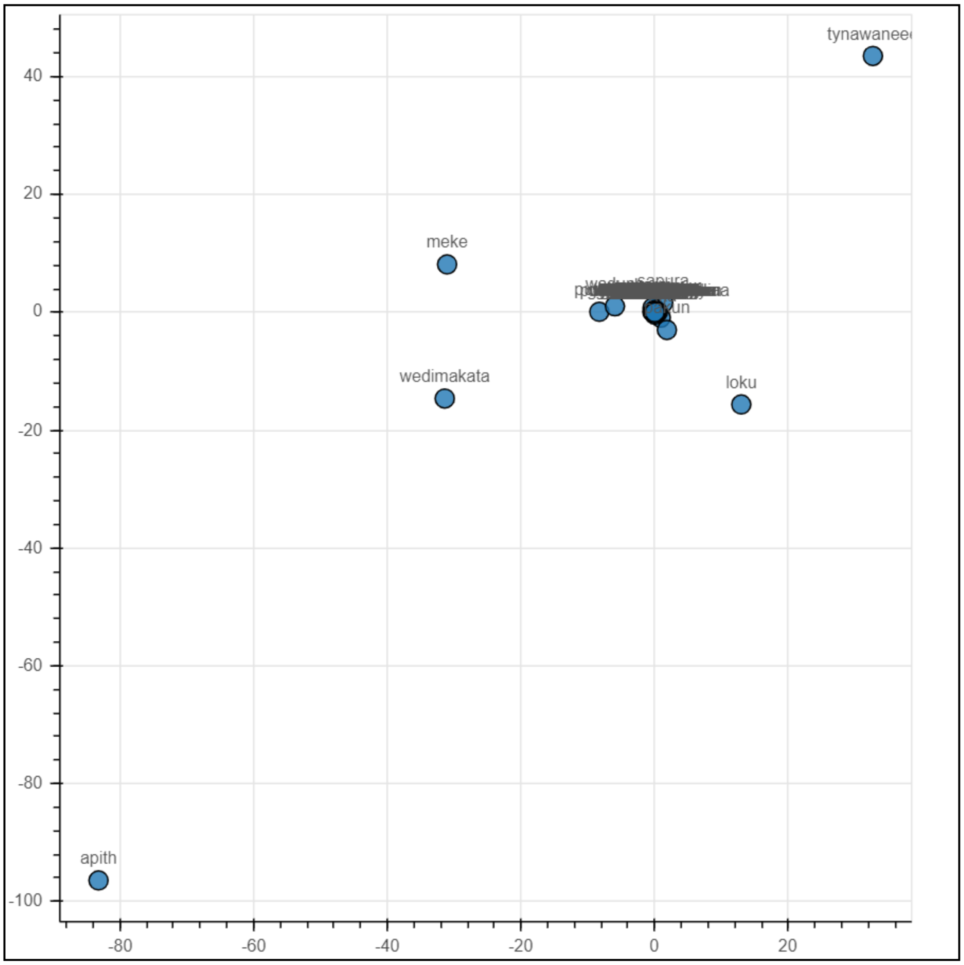


Figure 5: Word3Vec representation of Sinhala words in 2D space

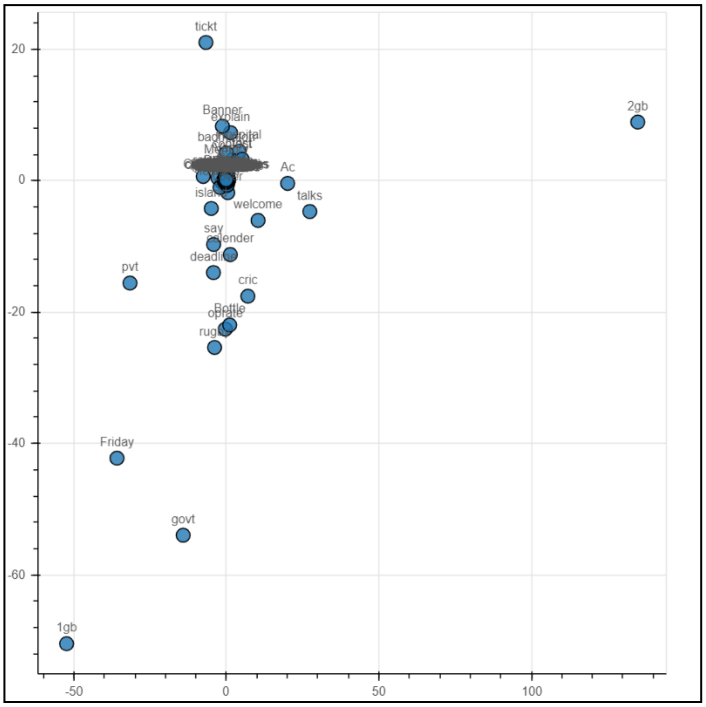


Figure 6: Word2Vec representation of English words in 2D space

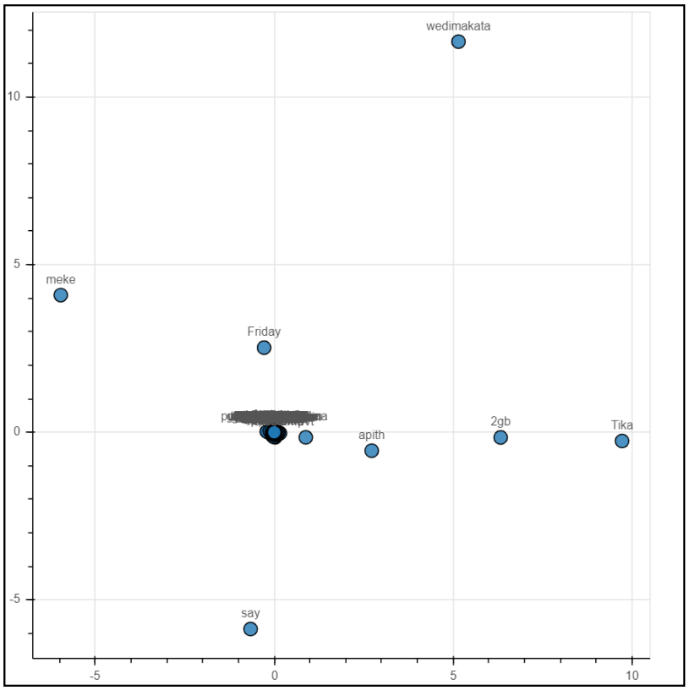


Figure 7: Word2vec representation of all tokens in 2D space

3.3 Code-mixed sentence classification

As stated earlier, first part of the language detection model is to create a classification model to successfully classify code-mixed sentences. That is, the assigning of labels depending on the mixture of languages within the sentence such as Singlish, English or code-mixed. Accordingly, the dataset generated in the level 1 annotation process was used in this study which consisted of 7,486 sentences with four labels as Singlish, code-mixed, English, and Unknown. Table 15 shows the features and the models tested in this process. Most of the features were selected based on similar studies and two new features were tested in this study which were not found in similar studies. That is word embedding features generated by two different Word2Vec models. In addition to the features, all the models were selected based on the classification problem. Further, each of the model was trained and evaluated with each of the feature and optimized to get the best performing model.

Table 15: Features and models tested for code-mixed sentence classification

Features	models
1. BOG (Bag of Words) 2. Word-level TF-IDF 3. TF-IDF 4. Character n-gram 5. Word embedding generated from one million web data (Word2Vec from Genisim) 6. Word embedding generated from the new data set (Word2Vec from Genism)	Non neural network-based modes 1. SVM 2. Random Forest 3. XGB Neural network-based models 4. Deep NN 5. CNN 6. Shallow NN 7. GRU 8. Recurrent CNN 9. LSTM 10. Bidirectional RNN

3.4 Sequence tagging

As the second part of the model creation, a sequence tagging model was tested on code-mixed dataset created within the second annotation phase. That is, the context clues were considered to label each token with a text based on its base language. Accordingly, a dataset with 1,900 sentences of Sinhala-English code-mixed data with four annotation tags were used to train each identified model. However, with the intension to detect only the language of the token, all the tokens with the label Name also categorized as Unknown resulting in a dataset with only three labels. Likewise, four models were selected based on the literature thereby resulting in three non-neural network models and one neural network model.

Table 16: Features and models tested for sequence tagging

Features	models
1. Character n-gram 2. Annotation of the left and right of the token 3. Capitalization 4. Whether the token is a digit or not	Non neural network-based modes 1. CRF 2. SVM 3. K-nearest neighbor Neural network-based models 4. LSTM

3.5 Experiment setup

All the experiments were done using a single EC2 instance in AWS and the hardware and software specifications of the setup is presented in table 17.

Table 17: Hardware and software configurations for experiment setup

Hardware	Software
• Instance type: m4.2xlarge • vCPU: 8	• OS: Linux • Core libraries:

• RAM: 32GB	▪ Python version: 3.6.4 ▪ Keras version: 2.2.5 ▪ TensorFlow version: 1.0 ▪ Numpy version: 1.16.0

4. RESULTS AND ANALYSIS

4.1 Results

4.1.1 Code-mixed data classification

As given in table 15, 10 machine learning models were tested with six features at a time. That is, each model accommodated all six features at a time and optimized it with a guided grid search algorithm with 10-fold cross validation to select the best model for model-feature combination. Likewise, sixty total model fittings (considering optimization as a single fit) were carried out to select the best model. On the other hand, the dataset had a class imbalance with the majority of the Singlish sentences and models tend to overfit to the majority class (Singlish) if used as it is. Thus, out of the well-known class balancing techniques (up sampling, down sampling and synthetic data generation), up sampling is used to balance the dataset. This is because, the smallest class is having only 13 sentences and down sample will remove majority of the sentences if it is used, therefore, it is deemed inappropriate. On the other hand, synthetic data generation is also eliminated as it might generate new sentences which may be meaningless or meaningless words. Further, up sampled data did not perform well with Random forest and XBM models. Therefore, the original dataset with class imbalance was used with class weights (optimized using grid search). And also, for n-gram feature, multiple n-grams were tested, and bigram is selected with overall accuracy across all models.

Subsequently, Naïve Bayes model was considered as the baseline model to all the models and figure 8 illustrates the comparison of all models. Accordingly, XGB model is the best performing model with an accuracy of 92.1% with character n-gram features. Further, XGB model outperformed Naïve Bayes model with word-level TF-IDF features and with locally trained word embedding features. Furthermore, it is observed that Random Forest and Logistic Regression models achieved their highest accuracy with character n-gram feature (figure 9). Moreover, SVM gave a constant accuracy for all the features and it did not outperform the baseline Naïve Bayes with any of the features. Likewise, it is also visible that none of the neural network-based model has been able to outperform Naïve Bayes model. In particular, the LSTM and GRU are the best performing models in that category with character n-gram feature (figure 10). Thus, the character n-gram is the best feature for this dataset for most of the models. Further, the deep neural network model shows an accuracy rate worse than a random guess and it indicates a significantly lower accuracy for all the features other than n-grm TF-IDF and with pre-trained word embedding model which was 50% (simply a random guess).

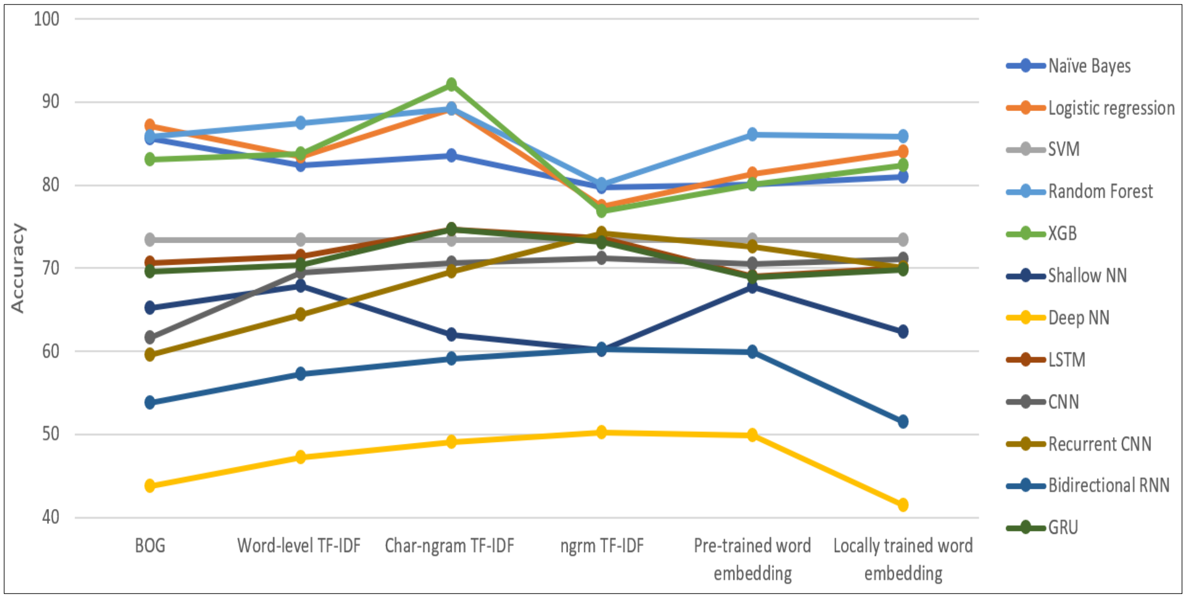


Figure 8: Accuracy comparison for all models

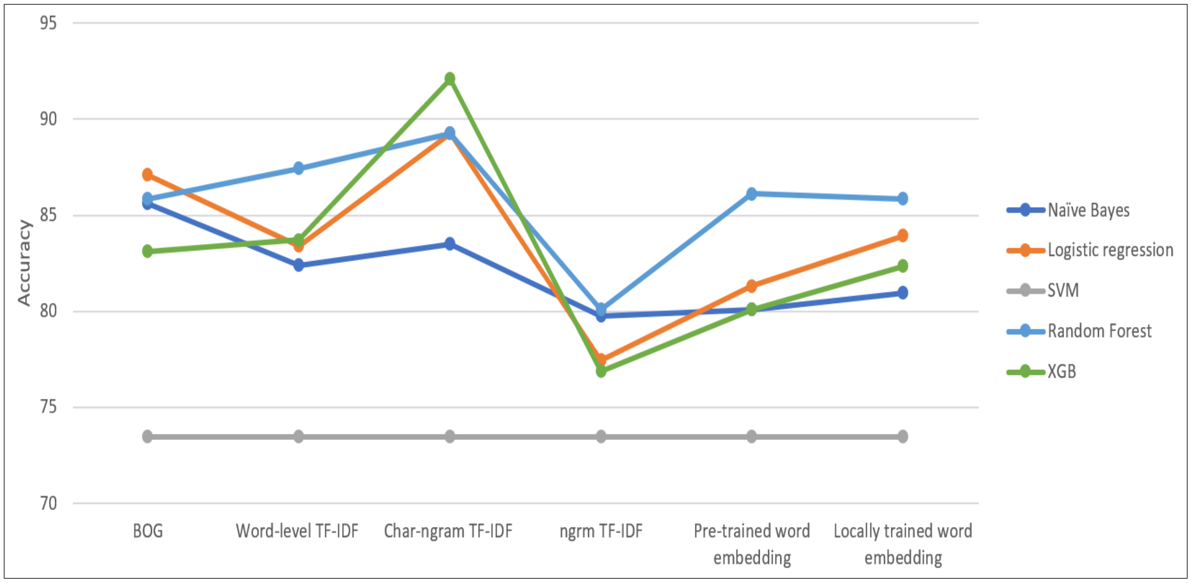


Figure 9: Accuracy comparison for non-neural network-based models

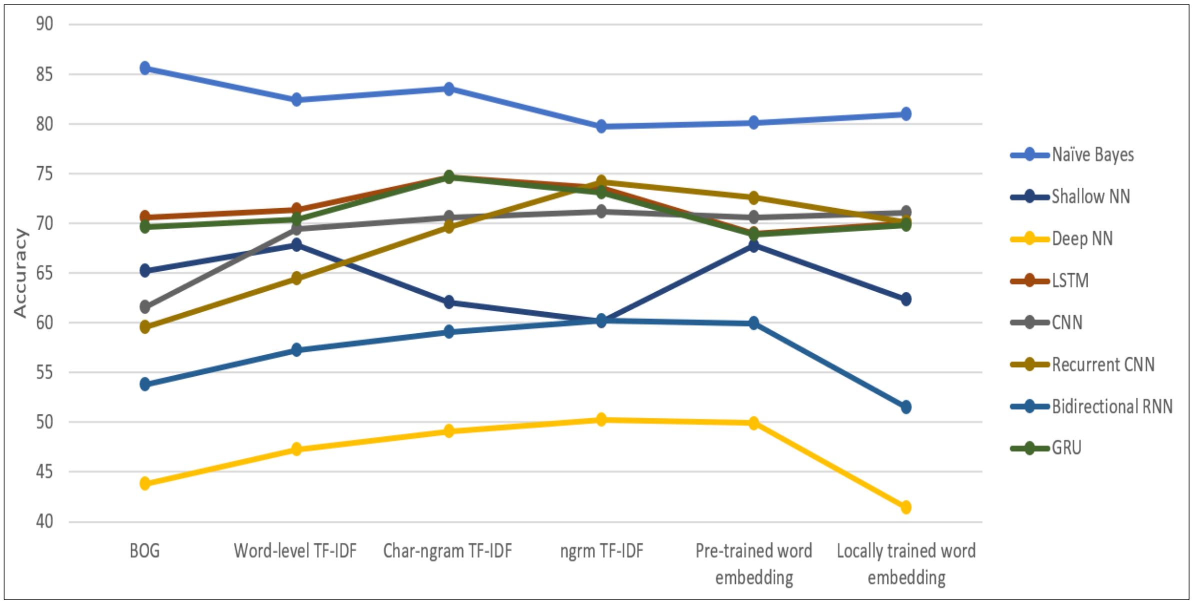


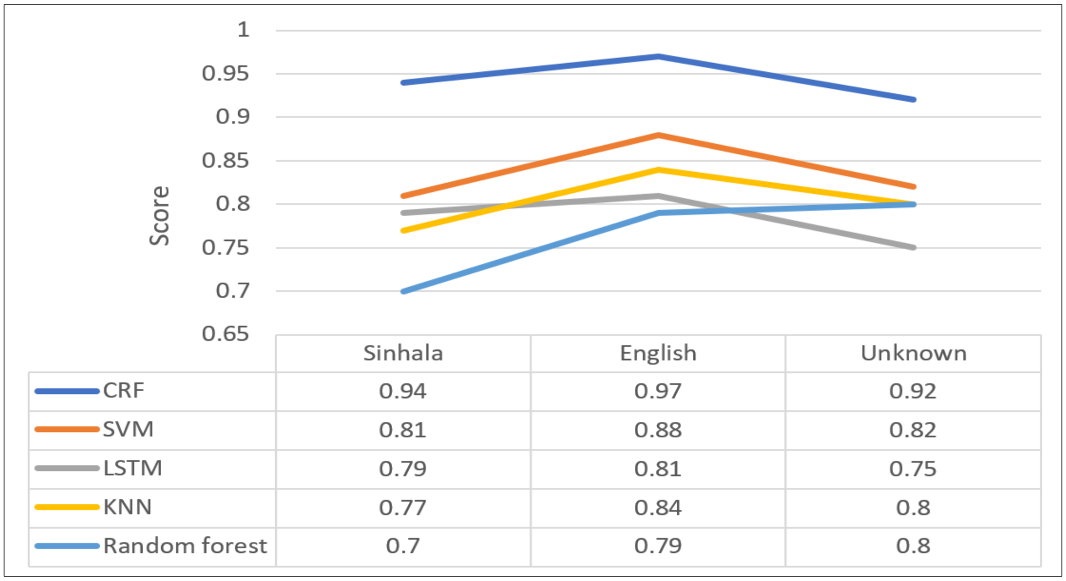
Figure 10: Accuracy comparison for all neural network models

4.1.2 Sequence tagging

According to table 16, four machine learning models were tested with four features in this scenario. Likewise, each model was trained and optimized with a guided grid search algorithm with 10-fold cross validation to select the best performing model for each case.

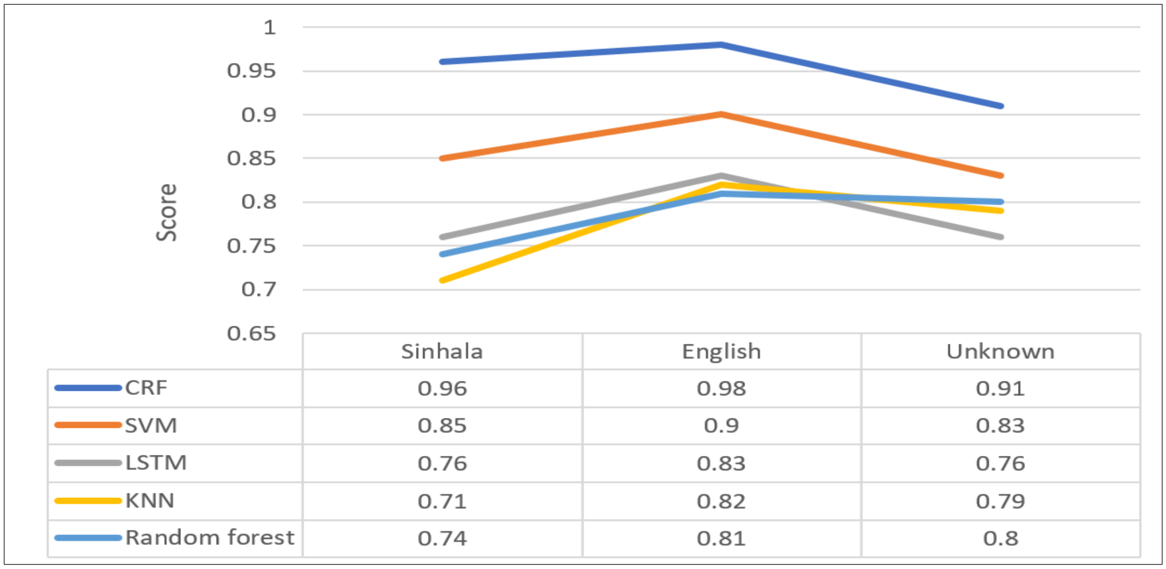
Likewise, 1,900 code-mixed sentences resulting in 11,795 tokens were used to train each model. On the other hand, the dataset was dominated by Sinhala tokens (72.64%). Since this a sequence tagging classification, none of the class balancing methods will fit into this scenario. Indeed, the usage of accuracy will also fail in this situation. That is, even with an overfitted model to Sinhala tokens will give an accuracy of 72.64%. Therefore, Precision, Recall and F1 scores were used as the performance matrix in this case to get more in-depth understanding on the model performance.

Likewise, CRF model outperformed all the other three models when compared to precision, recall and F1 score. It gave an overall precision of 0.95, recall of 0.94 and a 0.94 as F1 score (figure 11,12,13 and 14). That is, CRF model has been able to predict 95 of the actual languages of the tokens if 100 tokens (recall) were considered and out of the 100 tokens it predicted, 64 of them consisted the predicted language label (precision). On the other hand, it was also evident that CRF model has been able to accurately label many ambiguous words such as “10k”, “4i” and “ok”. Furthermore, it is noticeable that CRF model was performing with a larger margin compared to other models in all three performance matrices.



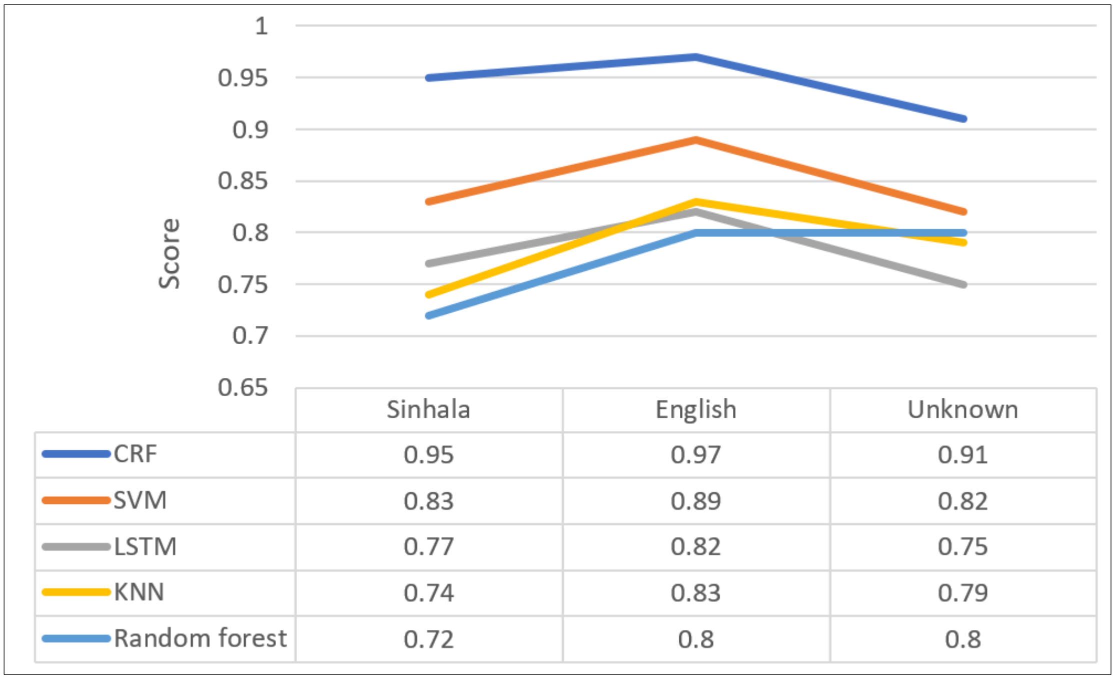
	Sinhala	English	Unknown
CRF	0.94	0.97	0.92
SVM	0.81	0.88	0.82
LSTM	0.79	0.81	0.75
KNN	0.77	0.84	0.8
Random forest	0.7	0.79	0.8

Figure 11: Comparison of precision scores



	Sinhala	English	Unknown
CRF	0.96	0.98	0.91
SVM	0.85	0.9	0.83
LSTM	0.76	0.83	0.76
KNN	0.71	0.82	0.79
Random forest	0.74	0.81	0.8

Figure 12: Comparison of recall scores



	Sinhala	English	Unknown
CRF	0.95	0.97	0.91
SVM	0.83	0.89	0.82
LSTM	0.77	0.82	0.75
KNN	0.74	0.83	0.79
Random forest	0.72	0.8	0.8

Figure 13: Comparison of F1 scores

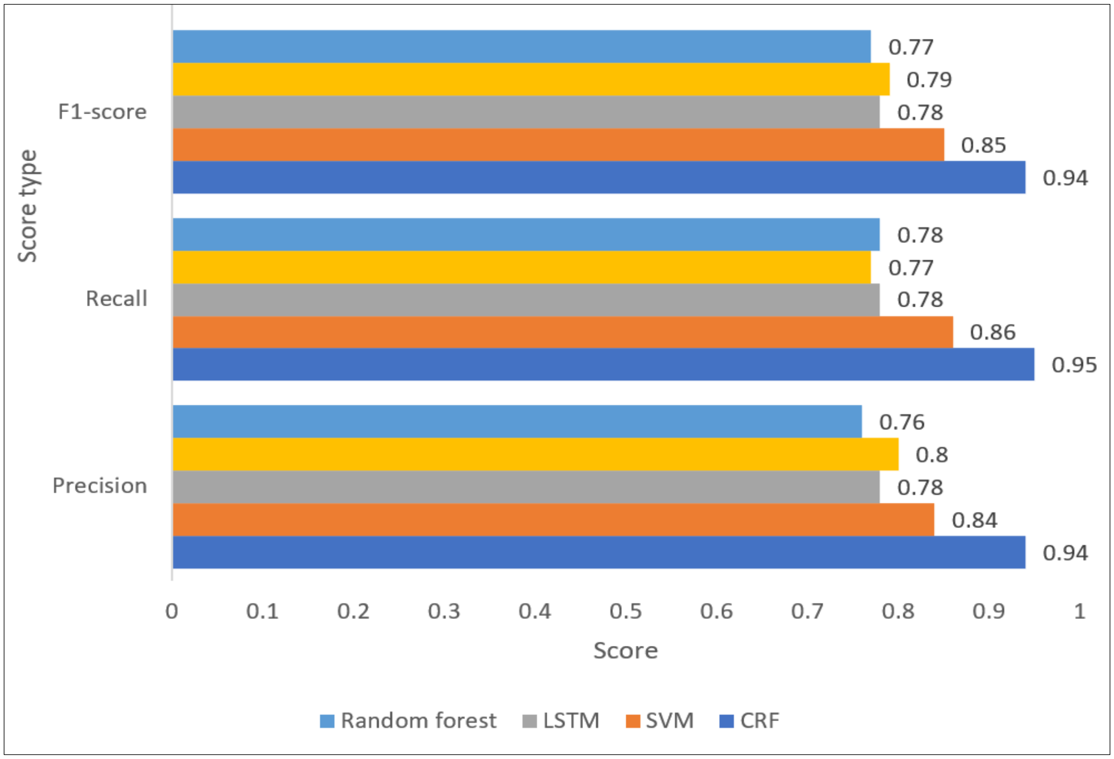


Figure 14: Comparison of average scores

4.2 Limitations and improvements

Accordingly, as per the results, it can be identified that none of the neural network models have outperformed the baseline model even after the optimizations. Further in the optimization, a guided optimization process was used to optimize each model and different model architectures were tested for each case. Moreover, within the model architecture selection, only selected architectures were used for optimization. For an example, 1D CNN, 2D CNN models with different number of convolution layers and pooling layers were used with manual selection and a wider search space was used with considering the time and memory concerns. Likewise, there can be a possibility where any other neural network-based model may perform well which has not been used in this study or the best hyper-parameters are out of the captured in the search space.

On the other hand, the size of the corpus was limited, and it seems not enough to train a deep learning model appropriately. Thus, another limitation of this model can be identified as limited number of samples used. However, within the practical scenario, Sinhala-English code-mixed data was not found in a larger scale and this was identified within the dataset analysis too. Despite of the significant use of code-mixed/Sinhala/Singlish data in various gossip sites by Sri Lankan users, majority of those texts consisted Unicode characters. In practice, most of the non-Unicode texts can be found in social media chat groups which were used as the data source for the new dataset. Nevertheless, even with that, it had only 1,900 code-mixed data out of 7,500 sentences (25.33%). Therefore, to generate a larger dataset of code-mixed data, representing the intended code-mixed sentences count, it is required to collect nearly four times large initial data. Yet, the appropriate level of annotation for sentence level and token level consumes a longer time. Indeed, within the current study it took nearly five to six months to prepare the full dataset along with the annotations at an acceptable level. Hence, with the timeline of the masters' course, generation of a larger data set was a tedious task. Yet, the author accepts that it is appropriate to analyze a larger data set to derive the conclusions.

Subsequently, in assessing the methodology all features used for the sequence tagging model creation were selected based on the literature reviewed. Thus, none of the upcoming stacking models have been used in the experiment. Indeed, the author picked these features and models from the literature reviewed to assess the compatibility of those models and features within the Sinhala-English code-mixed data scenario. On the other hand, upon reviewing the results of each study only the best performing models and features were selected based on their performance.

Further, the results of the current study revealed that character n-gram feature gave the highest accuracies for most of the models. Therefore, it would have been better if the study includes the "fastText" to represent words which is a popular word embedding technique along with Word2Vec in current context.

5. CONCLUSION

The purpose of the current study was to identify a machine learning model to recognize the base language of each token in a Sinhala-English code-mixed data context. Accordingly, since this was the first code-mixing data (written in Latin characters) analysis, the author developed a new dataset using Facebook data. Especially, this research focused on the code-mixed data written using Latin characters. Accordingly, a new dataset was created with all standard procedures and published to be used in future similar studies.

First, a code-mixed data classification model was created and when compared the results, XGB model outperformed with character n-gram features related to all the other tested models. Further, it was evident that tree-based models performed well in this scenario. In particular, the tree-based ensemble methods are the best performing models, among others. However, the neural network models were not able to outperform the baseline model. Even though LSTM models are well known for sequence predictions and compared to RNN models, it failed to outperform the baseline model. Thus, it can be concluded that a very low accuracies of neural-models/deep-learning model were generated due to the low number of data points. Indeed, the deep-learning models are well known to perform with larger datasets. Likewise, it can be concluded that there were not enough data points to train neural network models in this dataset because each model was tested with multiple configurations such as number of layers, number of nodes per layer, different batch sizes, with different epochs, and so on. Thus, the size of the dataset may have created implications to train the neural network models ineffectively.

Similarly, with manual inspection of the dataset, there were multiple occurrences of spelling mistakes and there were some scenarios where multiple words have been merged. Thus, token level features failed to catch such scenarios and character level n-grams have performed well in this case. Therefore, it can be concluded that character n-gram features are the best feature to be used in social media data analysis where spelling mistakes and merge of tokens are unavoidable.

Further, within the sequence labeling model, CRF has outperformed all the other models with a significant margin and this was presented within similar studies as well. Again, LSTM has not performed well in this case as well. In fact, it may be due to the lack of data points to train the model effectively.

6. FUTURE WORK

At first, in future studies, the dataset will be expanded to use multiple social media platform data such as Twitter, YouTube, and WhatsApp. In particular, this will lead to a data set with enriched variations in many social media platforms which can be analyzed for the behavior of each platform.

Further, during the current study the dataset contained hate words used by Sri Lankans. However, they were not omitted within the current study. Hence, those could be used for a hate speech analysis study. Likewise, a hate speech detection study will be conducted after enriching the dataset.

Furthermore, none of the stacked models were tested in this study. Therefore, in future studies combination of models such as CRF-LSTM, CNN-LSTM and deep LSTM models will be tested.

REFERENCES

[1] 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

[2] 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

[3] Blom, J.-P. and Gumperz, J. (1972), Social meaning in linguistic structures: code-switching in Norway, In J. J. Gumperz and D. Hymes (eds.), Directions in Sociolinguistics (pp. 407-34), New York: Holt, Rinehart and Winston

[4] Bokamba, E. G. (1989). Are there syntactic constraints on code-mixing?. World Englishes, 8(3), 277-292

[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] Boyd, D. M. & Ellison, N.B. (2007). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13.1: 210-30. Web

[7] Brown (1987). Principles of language learning and Teaching, London: prentice Hall

[8] Brown, Gillian. Yule, George. (1983). Discourse analysis. Cambridge: Cambridge University Press

[9] Cann, A. J., Dimitriou, K & Hooley. T (2011). Social media: A Guide for Researchers. Research Information Network, Web.

[10] Carter, S, et al. (2013) Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text. Language Resources and Evaluation, vol. 47, no.1, 195–215

[11] Chan, B. H. S. (1993), In search of the constraints and processes of code-mixing in Hong Kong Cantonese-English bilingualism, Research Report 33, Department of English, City Polytechnic of Hong Kong

[12] Chan, B. H. S. (2009). English in Hong Kong Cantopop: language choice, code-switching and genre. World Englishes, 28(1), 107-129

[13] Corder, S. (1973). Introducing applied linguistics

[14] Clyne, M. (1991), *Community languages: The Australian experience*, Cambridge: Cambridge University Press

[15] 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

[16] 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

[17] Faerch, C. and G. Kasper, (1983). Strategies in Inter-language Communication. Harlow: Longman

[18] Gibbons, J. (1987). Code-mixing and code choice: A Hong Kong case study (Vol. 27). Clevedon: Multilingual Matters

[19] Gumperz, J. J. 1982. Discourse Strategies. Cambridge: Cambridge University Press

[20] Heath, J. (1989). From code switching to borrowing: foreign and diglossic mixing in Moroccan Arabic (Vol. 9). Routledge

[21] Hidayat, T. (2008). An analysis of code switching used by facebookers

[22] 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).

[23] Iaia, P, L. (2016). Analysing English as a Lingua Franca in Video Games: Linguistic Features, Experiential and Functional Dimensions of Online and Scripted Interactions. Digital

[24] Lobov, W. (2001). Principals of Linguistic Change. Blackwell Publishers: University of Pennsylvania

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

[26] Montes-Alcalá, C. (2000). Attitudes towards oral and written codeswitching in Spanish-English bilingual youths. Research on Spanish in the US, 218-227

[27] Muysken, P., Díaz, C. P., & Muysken, P. C. (2000). Bilingual speech: A typology of code-mixing (Vol. 11). Cambridge University Press

[28] Myers-Scotton, C. (1988), Codeswitching as indexical of social negotiations, In M. Heller (ed.), Codeswitching (pp. 151-86), Berlin: Mouton de Gruyter. Social Motivations for Codeswitching, Oxford: Clarendon Press

[29] Myers-Scotton, C. and Jake, J. L. (1995), Matching Lemmas in a Bilingual Competence and Production Model, Linguistics 33: 981-1024

[30] Myers-Scotton, C. (1997). Duelling languages: Grammatical structure in codeswitching. Oxford University Press

[31] Nattinger, J.R (1992). Lexical Phrases Teaching. New York: Oxford University Press

[32] Pennington, M. C. (1996), Cross-language effects in biliteracy, Language and Education 10, 254-72.s

[33] Pennington, M. C. (1998). Language in Hong Kong at century's end (Vol. 1). Hong Kong University Press

[34] Poplack, S. (1980), Toward a typology of code-switching, Linguistics 18: 581-618

[35] 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

[36] Sankoff, D. and Poplack, S. (1981), A formal grammar for code-switching, Papers in Linguistics 14, 3-46

[37] Selinker, L. (1972). Interlanguage. IRAL-International Review of Applied Linguistics in Language Teaching, 10(1-4), 209-232

[38] Selinker, L. and Douglas, D. (1985). Wrestling with Context in Inter-Language Theory International Review of Applied Linguistic, 6,190-204

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

[40] Sridhar, S. N., & Sridhar, K. K. (1980). The syntax and psycholinguistics of bilingual code mixing. Canadian Journal of Psychology/Revue canadienne de psychologie, 34(4), 407

[41] Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge University Press

[42] Sternson and Schuman’s (ed), (1974). New Frontiers in Second Language Learning. Rowley Mass: Nebary House

[43] Tannen, D. (Ed.). (1982). Spoken and written language: Exploring orality and literacy (Vol. 32). ABLEX Publishing Corporation

[44] Trudgill, (1983). Sociolinguistics: An Introduction to Language and Society

[45] Voss, C. R., Tratz, S., Laoudi, J., & Briesch, D. M. (2014, May). Finding Romanized Arabic Dialect in Code-Mixed Tweets. In LREC (pp. 2249-2253)

[46] 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

[47] Wardhaugh, R. (1985). How Conversation Works. Basil Blackwell: Oxford

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

[49] Daly LE, Bourke GJ. Interpretation and Uses of Medical Statistics . 5th ed. Oxford, England: Blackwell Science;2000 .

[50] I. Smith and U. Thayasivam, "Sinhala-English Code-Mixed Data Analysis: A Review on Data Collection Process," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, pp. 1-6.

[51] I. Smith and U. Thayasivam, "Language Detection in Sinhala-English Code-mixed Data," 2019 International Conference on Asian Language Processing (IALP), Shanghai, Singapore, 2019, pp. 228-233.