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

DECLARATION i

ACKNOWLEDGMENT ii

ABSTRACT iii

List of tables vii

List of figures vii

1.0 INTRODUCTION 1

1.1 Language identification 1

1.2 Languages in social media 2

1.3 Derived languages found in social media 5

1.4 Language detection and its use cases 6

1.4.1 Business related usage 6

1.4.2 Non-business-related use cases 7

1.5 Scope of the study 7

2.0 LITERATURE REVIEW 8

2.1 Multi language learning 8

2.1.1 Language transfer 8

2.1.2 Transfer of training 9

2.1.3 Language learning strategy 9

2.1.4 Second language communication strategies 9

2.1.5 Overgeneralization of target language rules 10

2.2 Code-mixing 11

2.2.1 Motivations of code-mixing 13

2.2.2 Myers-Scotton's model 15
2.2.3 Code-mixing types found in social media 17
2.3 Related work 18
3. METHODOLOGY 26
3.1 Data collection 26
3.1.1 Annotations 27
3.2 Dataset analysis 35
3.2.1 Frequent words 35
3.2.2 Ambiguous words 36
3.3.3 Word2Vec representation 38
3.3 Code-mixed sentence classification 41
3.4 Sequence tagging 42
3.5 Experiment setup 42
4. RESULTS AND ANALYSIS 44
4.1 Results 44
4.1.1 Code-mixed data classification 44
4.1.2 Sequence tagging 46
4.2 Limitations and improvements 49
5. CONCLUSION 51
6. FUTURE WORK 53
REFERENCES 54

List of tables

Table 1: Natural language processing description in different languages 1
Table 2: Social media categories 3
Table 3 Language variation types found in social media 4
Table 4: Recent studies in automatic language detection since 2012 22
Table 5: Short form tokens and their probable complete token 27
Table 6: Tags used for level 1 annotation 28
Table 7: Languages mix after level 1 annotation 30
Table 8:Kappa values (Inter annotator agreement) for level 1 annotation 31
Table 9: Total language wise sentence count 31
Table 10: Summery of Level 2 annotation 34
Table 11: Unique token usage 35
Table 12: Frequent Sinhala words 36
Table 13: Frequent English words 36
Table 14: Ambiguous words 37
Table 15: Features and models tested for code-mixed sentence classification 41
Table 16: Features and models tested for sequence tagging 42
Table 17: Hardware and software configurations for experiment setup 42

List of figures

Figure 1: Sample of level 1 annotation process 30
Figure 2: Graphical representation of total language wise sentence count 32
Figure 3: Sample of Level 2 annotation sheet 33
Figure 4: Graphical representation of level 2 annotation statistics 35
Figure 5: Word3Vec representation of Sinhala words in 2D space 39
Figure 6: Word2Vec representation of English words in 2D space 40
Figure 7: Word2vec representation of all tokens in 2D space 40
Figure 8: Accuracy comparison for all models 45
Figure 9: Accuracy comparison for non-neural network-based models 46

Figure 10: Accuracy comparison for all neural network models 46
Figure 11: Comparison of precision scores 47
Figure 12: Comparison of recall scores 48
Figure 13: Comparison of F1 scores 48
Figure 14: Comparison of average scores 49