WORD LEVEL LANGUAGE IDENTIFICATION OF CODE-MIXING TEXT IN SOCIAL MEDIA USING NLP

Kasthuri Shanmugalingam

168287D

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

April 2019

WORD LEVEL LANGUAGE IDENTIFICATION OF CODE-MIXING TEXT IN SOCIAL MEDIA USING NLP

Kasthuri Shanmugalingam

168287D

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

April 2019

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a 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 person 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 thesis/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).

Name of Student

Kasthuri Shanmugalingam

Signature of Student: Date:

The above candidate has carried out research for the Master's Dissertation under my supervision.

Name of Supervisor

Dr. Sagara Sumathipala

Signature of Supervisor: Date:

Acknowledgements

I would like to make this opportunity to express my sincere gratitude to my supervisor Dr. Sagara Sumathipala for his valuable guidance extended throughout the research. This research would not have been completed to success without his immense support and guidance. Further, I would like to thank prof. Asoka Karunananda to gave valuable guidance to the documentation of the work during the lectures. And I would like to thank all academic staff of the Department of Computational Mathematics to gave sufficient knowledge to complete this research.

I wish to express my sincere thanks to my family members and colleagues who stood beside me whenever I need and helped me always with support, advice, and encouragement to complete my research work successfully.

Abstract

Automatic analyzing and extracting useful information from the noisy social media content are currently getting more attention from the research community. Recent days people easily mixing their native language along with the English language together to express their thoughts in social media, using the Unicode characters written in Roman Scripts. Thus these types of noisy code-mixed text are characterized by a high percentage of spelling mistakes with phonetic typing, wordplay, creative spelling, abbreviations, Meta tags, and so on. Identification of languages at word level become as necessary part for analyzing the noisy content in social media. It would be used as an intimidate language identifier for chatbot application by using the native languages.

For this study used Tamil-English and Sinhala-English code-mixed text from social media. Natural Language Processing (NLP) and Machine Learning (ML) technologies used to identify the language tags at the word level. A novel approach proposed for this system implemented as machine learning classifier based on features such as Tamil Unicode characters in Roman scripts, dictionaries, double consonant, and term frequency used for Tamil-English code-mixed text and features such as Sinhala Unicode characters written in Roman scripts, dictionaries, and term frequency used for Sinhala-English code-mixed text.

Different machine learning classifiers such as Support Vector Machines (SVM), Naive Bayes, Logistic Regression, Random Forest and Decision Trees used in the model evaluation process. Ten-fold cross-validation used to evaluate the performance based on language tags at the word level. Among that the highest accuracy of 89.46% was obtained in SVM classifier and 90.5% was obtained in Random Forest classifier for Tamil-English (Tanglish) and Sinhala-English (Singlish) code-mixed text respectively.

In the testing process of Tanglish model with SVM and Singlish model with Random Forest gave accuracy as 93.87% and 95.83% respectively for the testing unseen data. Tanglish model with SVM gave F-Measure for ‘tam’ and ‘eng’ tags were 0.965 and 0.894 respectively. Singlish model with Random Forest gave F-Measure for ‘sin’ and ‘eng’ tags were 0.975 and 0.929 respectively. So this the evidence that most of the times the Tanglish model with SVM and Singlish model with Random Forest predict the language labels correctly at word level.

Table of Contents