

**CROSS-LINGUAL DOCUMENT CLUSTERING FOR
SINHALA, TAMIL, AND ENGLISH USING
PRE-TRAINED MULTILINGUAL LANGUAGE
MODELS**

Malavarayar Vijayanandan Vithulan

(209390R)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2022

**CROSS-LINGUAL DOCUMENT CLUSTERING FOR
SINHALA, TAMIL, AND ENGLISH USING
PRE-TRAINED MULTILINGUAL LANGUAGE
MODELS**

Malavarayar Vijayanandan Vithulan

(209390R)

Thesis report submitted in partial fulfilment of the requirements for the degree
Master of Science in Computer Science specialisation in Data Science.

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2022

DECLARATION

I declare that this is my work. This dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or institute of higher learning. To the best of my knowledge and belief, it does not contain any previously published material written by another person except where the acknowledgment is made in the text.

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

Signature:

Date: 05/05/2022

The above candidate has researched the Master's thesis Dissertation under my supervision.

Name of the supervisor: Dr. Surangika Ranathunga

Signature of the supervisor:

Date:

ABSTRACT

Organising text articles into groups or clusters is known as document clustering. Documents that belong to a cluster are about the same subject. Document embeddings should be in the same embedding space for the cross-lingual document clustering, i.e., similar documents should have similar vectors. Obtaining document embedding for Tamil and Sinhala is feasible using models like Word2Vec or FastText, however, these embeddings are language specific, i.e., these will not be in the same vector space. Therefore, one cannot cluster documents across the languages using the language specific models. Pre-trained multilingual language models such as mBERT, XLM-R were introduced to solve this problem by transferring the knowledge from high resource languages to low resource languages.

This research is conducted to cluster Tamil, Sinhala and English news articles using XLM-R models. An adequate amount of collected documents were clustered, and the clustering techniques and performance were evaluated. This research produces a new baseline for cross-lingual clustering of Tamil, Sinhala, and English documents.

Keywords: Cross-lingual document clustering, Multilingual language models, XLM-R, mBERT, LASER, Knowledge distillation

ACKNOWLEDGEMENT

I want to express my deep and sincere gratitude to Dr. Surangika Ranathunga for guiding me in finding an exciting research topic and for continued support and encouragement. Her supervision immensely helped me in setting goals and engaging in the study.

I want to express my most incredible gratitude to the Department of Computer Science and Engineering, the University of Moratuwa, for providing the support to overcome this effort. Last but not least, my heartfelt gratitude goes to my parents and friends who supported me throughout this endeavour.

TABLE OF CONTENTS

DECLARATION	1
ABSTRACT	2
ACKNOWLEDGEMENT	3
TABLE OF CONTENTS	4
LIST OF FIGURES	6
LIST OF TABLES	6
LIST OF ABBREVIATIONS	7
1.0 INTRODUCTION	1
1.1 Research problem	2
1.2 Research objectives	2
2.0 LITERATURE REVIEW	3
2.1 Text representation	3
2.1.1 Classical methods	3
2.1.1.1 One hot encoding	3
2.1.1.2 Bag-of-Words (BoW)	3
2.1.1.3 Term Frequency (TF)	4
2.1.1.3 Term Frequency-Inverse Document Frequency (TF-IDF)	4
2.1.2 Word Embeddings	4
2.1.2.1 Word2Vec	4
2.1.2.2 Global Vectors (GloVe)	5
2.1.2.3 FastText	5
2.1.3 Contextual Embeddings	5
2.1.4 Transformer-based Pre-trained Language Models	6
2.1.4.1 GPT (OpenAI Transformer)	6
2.1.4.2 Bidirectional Encoder Representation from Transformers (BERT)	6
2.1.4.3 BART	7
2.1.4.4 Text-to Text Transfer Transformer (T5)	7
2.2 Multilingual Language Modelling (MLLM)	8
2.2.1 Architecture	8
2.2.2 Training Objective Functions	9
2.2.2.1 Parallel-Corpora based objective functions	9
2.2.2.2 Objective functions based on other parallel resources	11
2.2.3 Pre-training data and Languages	11
2.2.3.1 Curse of Multilinguality	11

2.2.4 MLLM Models	11
2.2.4.1 LASER	12
2.2.4.2 Multilingual BERT (mBERT)	12
2.2.4.3 XLM (Cross-lingual Language Model)	13
2.2.4.4 XLM-R (XLM-RoBERTa)	14
2.2.4.5 XLM-E (XLM-ELECTRA)	14
2.2.5 MLLM Benchmarks	15
2.2.6 Multilingual Knowledge Distillation	16
2.3 Document Clustering	17
2.3.1 Clustering approaches	17
3.0 RELATED WORKS	20
Vector alignment in multilingual word embedding	23
3.2 Summary	23
4.0. METHODOLOGY	25
4.1 System architecture	25
4.2 Models	25
4.2.1 Baseline	25
4.3 Clustering algorithms	26
4.4 Accuracy Algorithm	27
4.5 Optimal length finding algorithm	27
4.6 Input variations	27
4.7 Implementation	28
5.0 EVALUATION	29
5.1 Dataset	29
5.1.1 Data collection	29
5.1.2 Data quality	31
5.1.3 Inter annotator agreement (IAA)	31
5.2 Data description	31
5.2.1 Language breakdown	31
5.2.2 Categories breakdown	32
5.3 Results	32
5.3.1 XLM-R Base and One-Pass clustering	34
5.3.1.1 XLM-R Base and One-Pass clustering summary	35
5.3.2 XLM-R Base and K-Means clustering	36
5.3.3 XLM-R Large and One-Pass clustering	36
5.3.3.1 XLM-R Large and One-Pass clustering summary	37
5.3.3.2 Language-level performance breakdown	38

5.4 Discussion	39
5.4.1 Error analysis	42
6.0 CONCLUSION	44
7.0 REFERENCE	45

LIST OF FIGURES

Fig.1. Cross-lingual language model pre-training. This compares the difference between MLM and TLM
Fig.2 mBERT Accuracy of nearest neighbour translation for EN-DE, EN-RU and HI-UR
Fig.3. Overview of pre-training tasks in XLM-E
Fig.4. XTREME leaderboard summary as of September 25th 2021
Fig.5. Overview of knowledge distillation architecture
Fig.6. One-pass clustering algorithm logic
Fig.7. Parallel sentence retrieval accuracy after alignment in BERT
Fig.8. Zero-shot cross-lingual transfer result in XLM-R
Fig.9 System architecture
Fig.10. Categories breakdown in the collected data
Fig.11 Performance comparison of XLM-R models against the baseline
Fig.12 Performance comparison for KMeans and One Pass
Fig.13 Performance comparison between XLM-R Base and XLM-R Large models

LIST OF TABLES

Table 1. Comparison of popular language models
Table 2. Comparison of existing multilingual models
Table 3. XLM cross-lingual performance comparison
Table 4. XLM-R performance comparison
Table 5 Comparison of clustering algorithms used in document clustering
Table 6 Summary of approaches conducted in the researches
Table 7 Data sources where the data has been collected
Table 8 Language breakdown in the collected data
Table 9.1 A cross-lingual cluster with Ta and En that was formed using mBERT
Table 9.2 Correct cross-lingual cluster with Ta, Si and En that was formed using mBERT
Table 9.3 A correct cross-lingual cluster with Si and En that was formed using mBERT
Table 10. XLM-R base and One-Pass clustering accuracy summary

Table 11 Accuracy comparison against the content lengths in XLM-R Base
Table 12. XLM-R base and K-Means clustering accuracy summary
Table 13. Summary of XLM-R Large based clustering approaches
Table 14 Accuracy comparison against the content lengths in XLM-R Large
Table 15 A cross-lingual cluster with Ta and Si, that was formed using XLM-R Large
Table 16 A correct cross-lingual cluster with Ta, Si and En, that was formed using XLM-R Large
Table 17 Language level accuracy breakdown in XLM-R Large
Table 18 Summary of the accuracy scores among different approaches
Table 19 Language level performance breakdown with XLM-R Large and One-Pass
Table 20 Erroneous cluster formed due to order of the documents
Table 21 Erroneous cluster with shared word
Table 22 Erroneous cluster with shared context

LIST OF ABBREVIATIONS

Abbreviation	Description
NLP	Natural Language Processing
NER	Named-entity recognition
MLLM	Multilingual Language Models
mBERT	Multilingual BERT