A Singlish Supported Post Recommendation Approach for Social Media

Umesha Sandamini, Kusal Rathnakumara, Pasan Pramuditha, Madushani Dissanayake, Disni Sriyaratna, Hansi De Silva and Dharshana Kasthurirathna
Faculty of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Keywords: Singlish, Post Recommendation, Language Identification, Transliteration, Social Media.

Abstract: Social media is an attractive means of communication which people used to exchange information. Post recommendation eliminates the overflooding of information in social media to the users' news feed by suggesting the best matching information based on users' preference that in return increase the usability. Social media users use different languages and their variations where most of the Sri Lankan users are accustomed to use Sinhala and Romanized Sinhala. However, post recommendation approaches used in current social media applications do not cater to code-mixed text. Therefore, this paper proposes a novel post recommendation approach that supports Singlish. The study is separated into two major components as language identification and transliteration, and post recommendation. In this study, script identification was performed using regular expressions while a Naïve Bayes classification model that accomplished 97% of accuracy was employed for language identification of Romanized text. Transliteration of Singlish to Sinhala was conducted using a character level seq2seq BLSTM model with a BLEU score of 0.94. Furthermore, Google translation API and YAKE were used for Sinhala-English translation and keyword extraction respectively. Post recommendation model utilized a combination of rule-based and CF techniques that accomplished the RMSE of 0.2971 and MAE of 0.2304.

1 INTRODUCTION

Social media is a web-based or application-based online platform that behaves as a means of communication by facilitating the users to share their content. People around the world including Sri Lankans, use different social media platforms to share their thoughts. Among them, YouTube, Facebook, Instagram, Pinterest, WhatsApp can be identified as popular social media applications. According to the report 'Social media use in 2021' by the Pew Research Centre, YouTube and Facebook have been identified as the widely used platforms by American users ("Social Media Use in 2021," 2021) Figure 1. Furthermore, the report published by the same company in 2018 stated that 31% of the teenagers possess positive thoughts on social media based on the benefits gained such as the ability to connect with people, convenient and speedy access to news and information, the entertainment achieves through surfing, etc. (Anderson & Jiang, 2018).

People tend to use a variety of languages in social media. We conducted a survey by giving a questionnaire to Sri Lankan social media users. There, it was identified that most of the participants use Sinhala and Sinhala-English code-mixed texts rather than English to express their thoughts and information Figure 2. Sinhala-English code-mixing that is applied by end-users have different variations like phonetic typing of Sinhala words in Latin script (Romanized Sinhala), use of Sinhala and English words interchangeably in the same text, etc. Among them, phonetic typing of Sinhala words in Latin script is considered in this study.

Increasing the popularity of social media attracts more users towards them causing information flooding which in return reduces the user experience of the applications. As a solution to this issue, recommendation algorithms can be used to suggest information according to the users' preferences. Recommendation systems are categorized into demographic-based, content-based, and Collaborative Filtering (CF). Among them, CF and

Sandamini, U., Rathnakumara, K., Pramuditha, P., Dissanayake, M., Sriyaratna, D., De Silva, H. and Kasthurirathna, D.
A Singlish Supported Post Recommendation Approach for Social Media.
DOI: 10.5220/0010829700003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 412-419
ISBN: 978-989-758-547-0; ISSN: 2184-433X

content-based are the most often used filtering methods. Content-based filtering uses features, preferences, and similarity of content instead of user similarity. Similar users are defined as active users in CF (Fayyaz et al., 2020). In this study, a hybrid post recommendation approach based on the opinions and previously viewed posts of likely-minded persons is proposed.

Several text data pattern analysis techniques are used for recommendation purposes in social media to identify the interest of users. Extracted data in the analysis is accustomed to make a meaningful recommendation. Traditional approaches of recommendation systems generally utilize user-user, user-tag, and user-item relationships (Gao & Wolohan, 2016). Furthermore, the majority of the recommendation techniques don't support to Sinhala and Singlish languages used in social media networks. In the proposed approach posts are recommended based on users' interest, post view history, and post keywords. These parameters are used in the recommendation process to obtain the output that suits users' preferences. The recommendation system detects users' scenarios and provides a flexible approach to get prominent output.

In this study, a Sinhala and Singlish-supported post recommendation approach for social media is proposed which is discussed under two major components i.e. (1) Language Identification and Transliteration, (2) Personalized Post Recommendation. This paper includes a comparison of different methods and techniques that can be used for the subsections of the post recommendation approach proposed.

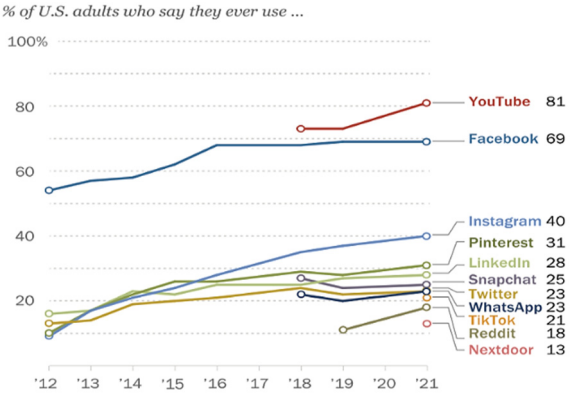

Figure 1: The percentages of social media usage by American adults. ("Social Media Use in 2021," 2021).

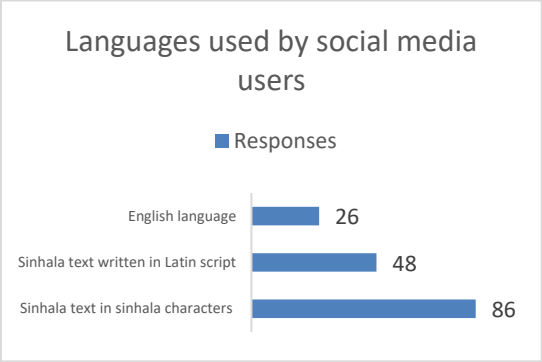

Figure 2: Summary of the survey conducted.

2 RELATED WORKS

In this section, existing works related to the two main modules of the proposed approach are reviewed.

2.1 Language Identification and Transliteration

A number of research have been carried out for language identification and transliteration of the code-mixed text to monolingual languages. Accordingly, a language identification model to classify English and Bengali subsequences in a code-mixed sentence using Long short-term memory (LSTM) has been conducted (Mahata et al., 2019). Patel and Parikh have used a Naïve Bayes approach along with a dictionary lookup to identify the respective language from the Gujarati-English code-mixed data (Patel & Parikh, 2020). However, both of these studies have not targeted the Sinhala and English code-mixed text.

Moreover, different approaches have been introduced by researchers for the conversion of code-mixed text to a monolingual language which is referred to as transliteration. It is transforming the words to the other language script with most matching letters rather than focusing on the sound of pronunciation. There are two ways of transliteration. i.e. (1) forward transliteration, (2) back-transliteration. The use of other foreign scripts to represent native text is known as forward transliteration whereas back transliteration is vice versa. There are three methods of performing transliteration. Automated Mapping of text that is written in one script to another based on language-specific rules is called transliteration generation. Transliteration mining is searching and extracting relevant pairs that are written in two scripts in

resources including parallel corpus, web, comparable corpus. The combination of both transliteration generation and mining is known as the hybrid or fusion approach. In this study, transliteration generation was applied. (Prabhakar & Pal, 2018)

In (Mahata et al., 2019) English segments have been translated using a character-based sequence-to-sequence(seq2seq) model with attention mechanism and back transliteration of Romanized Bengali segments to Devanagari has been performed using a seq2seq character-based model. Although, the model has shown a testing accuracy of 48.2% which is comparatively low. Furthermore, a language model has been built to solve the grammatical errors that arise when combining translated and back transliterated subsequences of text. Similarly, a dictionary lookup-based approach for transliteration of Gujarati text written in Roman script to the native script has been proposed. The variations that are not contained in the dictionary have been handled using Google Translation API (Patel & Parikh, 2020). As Sinhala is a low-resource language, it is difficult to find or develop corpora with a sufficient number of texts to obtain a good accuracy for language identification using this approach. Hence, this method is not much applicable in the Sinhala-English context.

Some studies have been conducted in the last few years for Sinhala-English code-mixed text. However, the number of studies conducted in this area is limited. In (Smith & Thayasivam, 2019) and (Shanmugalingam & Sumathipala, 2019) two different approaches for language identification of Sinhala-English code-mixed text have been introduced. Sentence level and word level annotation have been performed in (Smith & Thayasivam, 2019). The highest accuracy of 92.1% for language identification has been gained for the XGB model with bigram surpassing Support Vector Machine (SVM) and neural network models. Similarly, Shanmugalingam and Sumathipala have introduced another approach for word-level language classification using machine learning models where 90.5% accuracy has been achieved for the Random Forest (RF) classifier (Shanmugalingam & Sumathipala, 2019). Nevertheless, almost all the existing research on language identification of Singlish texts has targeted detecting Romanized Sinhala text when mixed up with English but none of them focused on identifying Singlish texts when mixed up with both pure English words and Sinhala words written in Sinhala Unicode.

Different approaches have been proposed for transliteration of Singlish to Sinhala in recent years. Liwera and Ranathunga have conducted a study applying a combination of trigram and rule-based approach that acquired a 77% of accuracy (Liwera & Ranathunga, 2021). In addition, a study for sentimental analysis of Sinhala-English code-mixed text has been carried out in which transliteration of Singlish to Sinhala was performed using a Singlish to Sinhala dictionary. The Singlish words identified in the comments have been replaced by the Sinhala words in the created dictionary. This approach has gained an accuracy of 72% (Sentimental Analysis of Comments in Social Media in Sinhala - English Code-Mixed Language Using Supervised Learning Techniques, 2020). However, in the rule-based approach mapping of English characters to Sinhala Unicode has to be performed manually and it requires a better understanding and knowledge of both languages. As there are multiple matches, the manually selected pair would not be the best possible candidate. De Silva has utilized an encoder-decoder LSTM model for Singlish to Sinhala transliteration with the use of a parallel corpus of 6000 Singlish words and received an accuracy of 40%. Although, the accuracy of the model was comparatively low (de Silva, 2019).

2.2 Recommendation System

Improvement of the social media recommender system decides the usability of users' activity. It has been stated that an ensemble model developed by merging the K-nearest neighbor and Naïve Bayes can be used to achieve a majority rating for a recommender system. In addition, fuzzy c means, and SVM have been used in this work to evaluate the overall average accuracy(Fayyaz et al., 2020). Furthermore, in social media networks, recommendation models should be able to recognize users' dynamic information. An algorithm has been suggested by using time factors, social relationships, response information, and geolocation which has been capable of mitigating the drawbacks of the traditional CF algorithm. The proposed algorithm has recognized users' behavioral patterns in order to recommend content in social media networks while maintaining a high-level efficiency (Cheng et al., 2015).

Furthermore, in (Amato et al., 2019) it has been claimed that most of the latest generation recommendation systems have been built of pre-filtering modules and advocated a user-centered approach for recommendations. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) remained at 0.94 and 1.6 respectively in the model they have proposed.

In summary, a smaller number of research has been carried out for Sinhala-English code-mixed text and only a few have been conducted on both language identification and transliteration. Furthermore, no studies have been conducted in this domain considering posts written in Sinhala, Singlish, or English languages as input to the system and providing personalized posts recommendation to the users regardless of their language. In this study, a novel approach for post recommendation in social media is proposed as a solution for the identified shortcomings of existing research discussed above.

3 METHODOLOGY

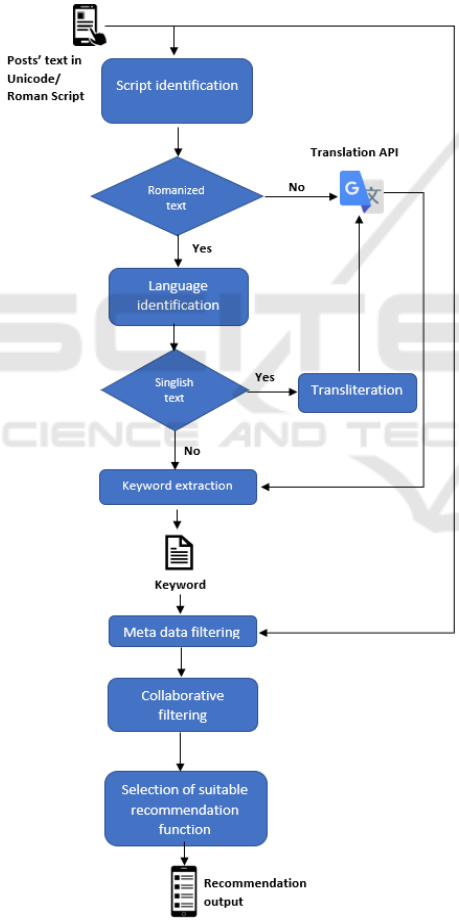

Figure 3: System diagram.

The proposed post recommendation approach includes two major components, i.e. (1) Language Identification and Transliteration, (2) Personalized Post Recommendation. As shown in Figure 3, once a post is input into the system, the text of the post is passed through the language identification and transliteration component. In there, the script in which the text is written identified as 'Unicode' or 'Roman', and the respective language of the text which is in Roman script is recognized as 'Singlish' or 'English'. Afterward, if the text is identified as Singlish it is transliterated to Sinhala. Subsequently, the result of transliteration or original text in Sinhala is translated to English and extract the core idea of each post. In this study, only the monolingual texts either in Sinhala, Romanized Sinhala, or English are considered. Keywords as the output of language identification and transliteration component are used for the personalized post recommendation component along with users' post view history, the content of the post, similar users, and rank. The hybrid recommendation model selects the most suitable recommendation method to feed the users with posts.

3.1 Language Identification and Transliteration

The language identification and transliteration module is further divided into 5 subsections as, (1) Script Identification, (2) Language Identification, (3) Transliteration, (4) Translation, and (5) Keyword Extraction.

3.1.1 Datasets

Two datasets were created using the text extracted from publicly available Facebook pages and groups and the corpus for transliteration was enhanced by combining the training data in the Dakshina dataset. In order to utilize for the language identification of posts, a dataset of 1853 samples was created. The corpus contained text written in either Singlish or English. Each sample of the corpus was labeled as 'Sin' or 'Eng' based on the respective language. A parallel corpus with 27393 unique Singlish words and their respective Sinhala words was created from 7110 texts gathered from the Dakshina dataset along with the author collected data from Facebook. It was pre-processed by re-moving all non-alphabetical characters and converted to lower case. Furthermore, the parallel corpus was cleaned by removing samples if its Singlish texts contain any non-roman characters or Sinhala text contains any roman characters. The dataset for language identification was pre-processed by removing URLs, emails, all non-alphabetical and non-numerical characters, wordplays, accented characters and then converted to the lowercase. Removing of wordplays signifies the trimming of words so as

not to contain the same character repeating consecutively more than twice (e.g. if the word was written as ‘owwwww’ trimmed to ‘oww’).

3.1.2 Script Identification

URLs, emails, accented characters, all non-alphabetical characters including punctuations, digits of the input text were removed as a pre-processing step before the script identification phase began. In the script identification phase, the script of the monolingual texts was detected as Sinhala Unicode or Latin script by making use of regular expressions. The Romanized Sinhala and pure English text were grouped under the Latin script.

3.1.3 Language Identification

In this section, the respective language of the Romanized text was identified as Singlish or English. Document-level classification of the monolingual Romanized text was performed. TF-IDF tokenized n-grams and Doc2Vec embedded texts were compared by classification using multiple machine learning algorithms i.e. SVM, Naïve Bayes, RF, Decision Tree, and Logistic Regression. Grid Search technique was used to recognize the optimized values for hyper-parameters.

3.1.4 Transliteration

Transliteration is required in transforming the user input to a common language for easy analysis. Initially, a rule-based approach was proposed to be used for transliteration of identified Singlish texts to Sinhala but due to multiple phonetic mapping between Sinhala and Singlish, a deep learning approach was decided to be used. Seq2seq model was used for the purpose due to its capability of self-learning the rules. The seq2seq model comprises an encoder and a decoder. Both of them are sequential-based models (Recurrent Neural Network (RNN), LSTM, BLSTM) which have been introduced to overcome the issue of RNN of not producing output with arbitrary length (Shah, 2020)(Singh, 2020). Two vocabularies were created for Singlish and Sinhala and converted the text in corpora to sequences of integers using the Keras Tokenizer class. Singlish was considered as the input language while Sinhala became the target language. Initially, both the input of the encoder and decoder was passed through an embedding layer. The encoder was built using a Bidirectional Long short-term memory (BLSTM) layer while a decoder with an LSTM layer was used. In this study, a character level approach was used where the encoder-decoder model learns the input character provided and then predicts the next character based on the probability of the likelihood of occurring that character. The hidden states of the encoder are passed to the decoder as its initial state. During the training process, the decoder predicts one character at each timestamp using teacher forcing where the actual output of the previous timestamp is fed to the LSTM cell instead of the predicted output (Shah, 2020)(Shirsath, 2021). The prediction was performed without teacher forcing. In predicting Sinhala words, it was made sure to input only the Romanized text to the seq2seq model. Eventually, numerical values and other non-Romanized characters in the input Singlish text were attached to the same location of the resultant Sinhala text of the transliteration phase.

3.1.5 Translation

In order to perform Sinhala to English translation, Google Translation API was used. Original user inputs in Sinhala or the output of the transliteration phase were subjected to translation.

3.1.6 Keyword Extraction

Yet Another Keyword Extractor (YAKE), an unsupervised approach that extracts keywords regardless of the size, language, and domain of the text (Campos et al., 2020) was compared with the KeyBERT, a mechanism that utilizes BERT embedding and Cosine Similarity for keyword extraction.

3.2 Recommendation System

The hybrid recommendation module of the proposed approach categorized the information posted in social media based on keywords that were generated in the language identification and transliteration module. The similarity of posts was calculated based on the keywords. The recommendation model examined the user's view history to detect and evaluate the user's interest by utilizing temporal entropy to output posts that are personalized to the user.

The most accurate and appropriate method was chosen for the post recommendation by experimenting and testing with categorization ranking and metadata-based filtering.

3.2.1 Categorization and Ranking

Category mapping function analyzed dataset to detect keywords of the post content by applying TfidfVectorizer. The TF-IDF is utilized for feature extraction text in Natural Language Processing (NLP). In other

words, the system counts the number of times each word appears in a text, weighing the value of each word, and calculating a score for the document. The posts vectors that lay in the extent of proceeding angle computed by the cosine similarity calculation were input to the category similarity ranking.

3.2.2 Metadata based Filtering

The model combined the raw information with the keywords provided in the language identification and transliteration module to create a new parameter that serves as an input to the TfidfVectorizer. The feature vectors built using the TfidfVectorizer were input to the Cosine Similarity Calculation.

3.2.3 Collaborative Filtering (CF)

The matrix preferences were analyzed using the dataset built of posts, post view history of users, and users' profile information. The data frame created with the dataset was used to recommend posts for a particular user. The centralized CF-based recommendation predicted missing post preference ranking based on similar users and user engagement. Subsequently, post correlation was computed by spearman's correlation coefficient for posts and users.

3.2.4 Personalized Recommendation

The proposed model provided personalized post recommendations considering different aspects including,

- New users of the system are suggested by the top-ranked and viral posts published by the existing users.
- The current users are recommended with the top-ranked post that is viewed by similar users.
- Recommend posts considering the engagement of similar users and the similarity of the metadata of their posts.

which in return improves the quality of the recommendation process

4 RESULTS AND DISCUSSION

The main intention of this study was to develop a post recommendation system that supports Sinhala, Singlish, and English texts. The overall system was comprised of two major components as (1) Language Identification and Transliteration (2) Personalized Post Recommendation.

The language identification models were evaluated using accuracy, precision, F1 score, and confusion metric. According to Table 1 three models achieved the highest accuracy of 0.97 i.e. 1-gram based Naïve Bayes model, SVM, and RF model with Doc2Vec document embedding. However, it was identified using the confusion matrix that the Naïve Bayes model worked well for both Singlish and English rather than the other two models with 0.97 accuracy. It can be observed that the accuracies of the models that utilized 3-gram and 5-gram were less and overall models with Doc2Vec document embedding provided more accurate results. Taking the results of the language identification phase into consideration Naïve Bayes model that made use of word unigrams was selected as the final model for Singlish and English language detection.

Table 1: Summary of the accuracies of language identification models.

Feature extraction technique		NB	SVM	RF	DT
TF- IDF	1-gram	0.97	0.96	0.96	0.95
	3-gram	0.78	0.75	0.71	0.68
	5-gram	0.68	0.68	0.67	0.67
Doc2Vec		0.96	0.97	0.97	0.96

The seq2seq encoder-decoder model built for Singlish to Sinhala transliteration was evaluated using the Bilingual Evaluation Understudy (BLEU) score where Sinhala text in test data was compared with its predicted text. The model was able to achieve a BLEU score of 0.94 which outperformed the encoder-decoder LSTM model proposed in the study (de Silva, 2019). The training and validation accuracies of the model were 0.98 and 0.89 respectively.

Table 2: Output of language identification and transliteration component.

Singlish Text	Ihala wurthiya nipunathawak sahitha madya wurthikayin godanagima aramunayi
Predicted Sinhala text	ඉහල වෘත්තිය නිපුණතාවක් සහිත මධ්‍ය වෘතිකයින් ගොඩනැගීම ආරමුණයි
Actual Sinhala Text	ඉහල වෘත්තිය නිපුණතාවක් සහිත මාධ්‍ය වෘත්තිකයින් ගොඩනැගීම අරමුණයි
English Translation of the system	Aims to build middle class professionals with high professionalism
Actual English Translation	The aim is to build media professionals with high professionalism
Keywords	high professionalism, build middle, middle class, class professionals, Aims

As shown in Table 2, the language identification and transliteration component provided keywords when a new post is input. However, due to the vast variation in the mapping of Singlish to Sinhala, in some instances, there are a few differences in predicted transliterated text compared to the actual as shown in Table 3. Accordingly, 'madya' can be written as 'මධ්‍ය' which has the meaning of middle and 'මාධ්‍ය' which is 'media' in English language. 'ආරමුණයි', the predcited word does not possess any meaning but 'අරමුණයි' is the Sinhala translation for 'aim'. For the given sample input correct Sinhala mapping for 'madya' and 'aramuna' should be 'මාධ්‍ය' and 'අරමුණයි' respectively.

Table 3: Camparison of the Results of Table 2.

Singlish word	Predicted Sinhala word	Actual Sinhala word
Madya	මධ්‍ය	මාධ්‍ය
aramunai	ආරමුණයි	අරමුණයි

YAKE was decided to be used as the keyword extraction technique as it provided more relevant keywords when compared with the KeyBERT.

Table 4: Summary of the results in recommendation model evaluation.

	Mean	Std
RMSE	0.2971	0.0046
MAE	0.2304	0.0032

The hybrid recommendation module showed higher-ranked posts to the user, ordered by ranking level in ascending order. Based on the scenario of the user, the recommendation model suggests more pertinent posts. According to the results of the evaluation and output gained from the proposed approach, it was identified that a combined model for post recommendation is more applicable than a single model. Filtering based on metadata provided better outcomes than categorization and ranking. Cold start problem and inflexibility in applying side queries are the downsides of CF recommendation models. The proposed approach eliminated the above issues and provided a better post recommendation approach. According to Table 4, the mean of RMSE and MAE values for the post recommendation model are 0.2971 and 0.2304, respectively. Utilizing singular value decomposition, singular vectors from ungraded and scattered ranked post-view history records were factorized.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel post recommendation approach for social media that supports Sinhala, Singlish, and English languages. For this study, only the monolingual texts were considered. Mainly the work is divided into two major modules as language identification and transliteration, and post recommendation. The data for the study was collected by scraping publicly available Facebook pages and groups through a third-party scraping tool and the parallel corpus for transliteration was enhanced by combining with the Dakshina dataset.

Script of a particular post was detected with the use of regular expressions. The texts written in Latin script were input to the language identification phase where the language of the texts was identified as 'Singlish' or 'English'. A machine learning model that built of Naïve Bayes classifier and considered word unigram with 97% accuracy was used for the language identification. In order to transliterate Singlish text to Sinhala a character level seq2seq BLSTM model with a BLEU score of 0.94 was utilized. Translation of Sinhala texts to English was performed using Google Translation API. YAKE was employed as the keyword extraction mechanism as it provides accurate keywords regardless of the size of the text and its domain.

A combined approach of rule-based and CF-based was used for the post recommendation module. This study solves the downsides of the traditional recommendation models. The recommendation module was evaluated using RMSE and MAE. The validation was conducted using 5-fold cross-validation and received RMSE as 0.2971 and MAE as 0.2304.

Developing the parallel dataset was the hardest part of this study. Though the transliteration phase achieved good results, still there are some differences between the predicted and actual output due to multiple ways of Singlish mapping to Sinhala. This can be prevented to a certain extent by using a word level seq2seq model with enough size of Singlish and Sinhala parallel corpus. Moreover, the insufficiency of data caused less accuracy in the recommendation model. In the future, this post recommendation approach is expected to be improved to support Sinhala, Romanized Sinhala, and English mixed texts where the current study is only limited to the monolingual text. Further, the accuracy of the transliteration phase and recommendation module can be increased by using enhanced datasets and using the word-level transliteration approach.

REFERENCES

Amato, F., Moscato, V., Picariello, A., & Piccialli, F. (2019). SOS: A multimedia recommender System for Online Social networks. *Future Generation Computer Systems*, *93*, 914–923. https://doi.org/10.1016/j.future.2017.04.028

Anderson, M., & Jiang, J. (2018). Teens, social media & technology. In *Pew Research Center [Internet & American Life Project]*. http://publicservicesalliance.org/wp-content/uploads/2018/06/Teens-Social-Media-Technology-2018-PEW.pdf

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Cheng, J., Liu, Y., Zhang, H., Wu, X., & Chen, F. (2015). A new recommendation algorithm based on user's dynamic information in complex social network. *Mathematical Problems in Engineering*, *2015*. https://doi.org/10.1155/2015/281629

de Silva, A. D. (2019). *Singlish to Sinhala Converter using Machine Learning A dissertation submitted for the Degree of Master of University of Colombo School of Computing*.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences (Switzerland)*, *10*(21), 1–20. https://doi.org/10.3390/app10217748

Gao, Z., & Wolohan, J. (2016). *Fast NLP-based Pattern Matching in Real Time Tweet Recommendation*. 1–4. https://trec.nist.gov/pubs/trec26/papers/SOIC-RT.pdf

Liwera, W. M. P., & Ranathunga, L. (2021). *Combination of Trigram and Rule-based Model for Singlish to Sinhala Transliteration by Focusing Social Media Text*. 1–5. https://doi.org/10.1109/fiti52050.2020.9424880

Mahata, S. K., Mandal, S., Das, D., & Bandyopadhyay, S. (2019). Code-mixed to monolingual translation framework. *ACM International Conference Proceeding Series*, 30–35. https://doi.org/10.1145/3368567.3368579

Patel, D., & Parikh, R. (2020). Language Identification and Translation of English and Gujarati code-mixed data. *International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020*, 2014–2017. https://doi.org/10.1109/ic-ETITE47903.2020.410

Prabhakar, D. K., & Pal, S. (2018). Machine transliteration and transliterated text retrieval: a survey. *Sadhana - Academy Proceedings in Engineering Sciences*, *43*(6), 1–25. https://doi.org/10.1007/s12046-018-0828-8

Sentimental analysis of comments in social media in sinhala - english code-mixed language using supervised learning techniques. (2020). 2020.

Shah, D. (2020). *Machine translation with the seq2seq model: Different approaches*. https://towardsdatascience.com/machine-translation-with-the-seq2seq-model-different-approaches-f078081aaa37

Shanmugalingam, K., & Sumathipala, S. (2019). Language identification at word level in Sinhala-English code-mixed social media text. *Proceedings - IEEE International Research Conference on Smart Computing and Systems Engineering, SCSE 2019*, 113–118. https://doi.org/10.23919/SCSE.2019.8842795

Shirsath, A. (2021). *Neural Machine Translation Using Sequence to Sequence Model*. https://medium.com/geekculture/neural-machine-translation-using-sequence-to-sequence-model-164a5905bcd7

Singh, P. (2020). *A Simple Introduction to Sequence to Sequence Models*. https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models/

Smith, I., & Thayasivam, U. (2019). Sinhala-English Code-Mixed Data Analysis: A Review on Data Collection Process. *19th International Conference on Advances in ICT for Emerging Regions, ICTer 2019 - Proceedings*, 1–6. https://doi.org/10.1109/ICTer48817.2019.9023739

Social Media Use in 2021. (2021). In *Pew Research Center* (Issue April). https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/