Sinhala-English Code-Mixed Language Dataset with Sentiment Annotation

D. K. Uthpala
Department of Physical Science
Faculty of Applied Science
University of Vavuniya
Vavuniya, Sri Lanka
uthpaladil95@gmail.com

S. Thirukumaran
Department of Physical Science
Faculty of Applied Science
University of Vavuniya
Vavuniya, Sri Lanka
thirukumaran@vau.ac.lk

2024 4th International Conference on Advanced Research in Computing (ICARC) | 979-8-3503-8486-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICARC61713.2024.10499746

Abstract—**In communication, technology has been played a significant role in many ways, and it is an essential part for human life nowadays. The majority of people commonly speak two or more languages for making better communication in the regional level or worldwide. Code-mixing is a practice of mixing words from different languages in multilingual settings. In addition, there is a growing demand for code-mixed sentiment analysis of comments posted by users on social media. Systems are trained for data available in one language only and failed with the data in multiple languages, because of the complexity of mixed data at different levels. However, there are only very few code-mixed data are available to create a model. There are no resources available for Sinhala-English code-mixed language, and it is important for researchers to give attention on sentiment analysis using Sinhala-English mixed language. We present a sentiment-labeled corpus for sentiment analysis of code-mixed Sinhala-English text using comments from *YouTube*® videos. An annotation setup is used to label and create a Sinhala-English dataset for sentiment analysis and the comments are pre-processed to clean. The entire data set has been divided into three groups: neutral, negative, and positive. In order to demonstrate the insight of the dataset, this study employs five machine learning algorithms on a newly created Sinhala-English dataset and achieved significant accuracy.**

Keywords—Annotation, Code-mixed, Dataset, Sentiment Analysis, Sinhala-English

I. INTRODUCTION

Sentiment Analysis (SA) is a major application in the field of Natural Language Processing (NLP) and it is an active research area in Computer Science, concerned about analyzing sentiments or opinions making progress since the early '90s. It is a text analyzing method that detects polarity, either negative or positive within the text. This focuses on polarity but also emotion or feelings and even intentions. Sentiment Analysis is an opportunity to explore the mindset of the audience or customers and to study the state of the subject. SA has become an important research field regarding social media to make decisions. But the texts are usually used informally on social media, mostly on *YouTube*®, *Facebook*®, and *Twitter*®. Different users in terms of linguistic backgrounds and cultures, try to express their idea about the subject using Code-mixed languages to mark their impressions comfortably. Even though most languages have their scripts, code mixed language users usually use Roman scripts [1]. Our study focuses on SA in Sinhala-English (*Singlish*) which has no available sentiment annotated data for code mixed scenarios.

Sinhala is an Indo-Aryan language that is written using the Sinhala script. Having been in contact with a foreign language for over twenty decades, urban bilinguals in Sri Lanka have mixed a large number of vocabulary items in their daily conversations. Today, language fusion between Sinhala and English has become a common and natural phenomenon. Structural analysis reveals inclusion and duplication as key strategies used by respondents to mix languages. There are a lot of code-mixed data for Sinhala-English text on social media. Some Sociologists and linguistic researchers were involved in the Sinhala-English Sentiment analysis scenario. But still, NLP researchers in Sri Lanka do not touch the scenario significantly. Hence, a sentiment analysis dataset is presented for Sinhala-English mixed language.

II. LITERATURE REVIEW

The techniques for sentiment classification can be divided into machine learning, dictionary-based, and hybrid approaches. Linguistic features are used in machine learning approach. The Lexicon-based approach relies on a collection of known and pre-compiled sentiment terms. It is divided into dictionary-based approach [2] and corpus-based approach [3] using statistical or interpretive methods to find sentiment polarity. The hybrid approach [4], [13] combines the two approaches, machine learning and dictionary-based, and the sentiment lexicons take a major role in most of the methods.

There have been many applications and improvements to the Sentiment Analysis algorithms proposed over the past few years. Recently, considerable work and effort have been put into collecting resources for code-mixed text. However, there are limitations in number, size and availability for code-mixed datasets and dictionaries for sentiment analysis. Code-mixed data is a significant challenge in NLP, as it has different features than traditional standard languages [1], [5], [6], [12]. In a few code-mixed languages such as Malayalam-English [1], Hindi-English [6], [7], Chinese-English [8], Tamil-English [9], and Bengali-English [13] datasets are available for research. Several research projects on sentiment analysis for such languages are held because of the potential market of the commercial NLP applications as well as the fastest-growing research area in Computer Science.

The review paper [5] highlights the momentum in Hindi sentiment analysis compared to code-mixed Indian languages, citing scarcity of linguistic tools and annotated resources. It emphasizes challenges in code-mixed SA, including noise, language identification, and limited annotated datasets. The authors stress the need for multiple processing levels to effectively handle code-mixed data. Embarking on an exploration of SA challenges within the realm of code-mixed data, the study in [6] puts forth a comprehensive methodology. Addressing the unique complexities inherent in code-mixed text, the proposed approach encompasses language identification, word transliteration, sentiment score tagging, feature extraction, and supervised learning methods. Leveraging a social media dataset comprising 1200 Hindi and

979-8-3503-8486-4/24/$31.00 ©2024 IEEE

300 Marathi documents, including diverse forms of communication like chats, tweets, and *YouTube*® comments, the research underscores the effectiveness of machine learning algorithms. Notably, Naïve Bayes, Support Vector Machine (SVM), and Random Forest (RF) exhibit efficiency in sentiment classification for both languages, providing invaluable insights for future investigations in this dynamic research domain.

Building upon the challenges highlighted in code-mixed SA, the study by [9] presents, Corpus Creation for SA on Code-Mixed Tamil-English text is presented with substantial corpus for under-resourced code-mixed *Tanglish* (Tamil-English) with annotation for sentiment polarity using the approach described in [10]. It uses input features Term Frequency - Inverse Document Frequency (TF-IDF) values up to three grams to evaluate the dataset on the algorithms Logistic Regression (LR), SVM, Decision Tree (DT), RF, Multinomial Naive Bayes (MNB), Deep Max Entropy (DME), K-Nearest Neighbour (KNN), Character-level Deep Max Entropy (CDME), 1 Dimensional Convolution - Long Short-Term Memory (1DConv-LSTM), and Bidirectional Encoder Representations from Transformers (BERT)-Multilingual. It states that all the used machine learning algorithms performed poorly on the presented dataset. The algorithms LR, RF, and DT are fared comparatively better with the highest macro avg. for RF: 0.42. Additionally, [1] contributes a sentiment analysis dataset for code-mixed Malayalam-English, employing the same approach. To evaluate the performance of machine learning algorithms on this dataset, the study employs the same input features and algorithms as presented in [1]. Notably, the findings suggest that most machine learning algorithms, with the exception of SVM , demonstrate success in classifying all sentiment classes present in the dataset. The highest macro average, achieved by the BERT model, stands at 0.61, indicating its effectiveness in capturing sentiment patterns in the code-mixed text.

Introducing an innovative approach, [7] proposes Deep Learning Architecture for Code Mixed Text (DLACMT), a novel deep learning technique tailored for sentiment analysis of code-mixed Hindi-English text prevalent in social media. Through the fusion of character and word features, DLACMT achieves superior accuracy compared to baseline models, showcasing its resilience across various loss functions and optimizers. The study sheds light on the efficacy of DLACMT in decoding sentiments within multilingual social media discourse, contributing valuable insights to SA methodologies. Furthermore, [7] underscores the significance of crafting robust training corpora for word embeddings in code-mixed languages, opening avenues for future exploration in this domain.

In comparison to other research, the approach outlined in [11] uniquely addresses sentiment mining in code-mixed sentences involving English and four additional Indian languages (Tamil, Telugu, Hindi, and Bengali). The paper presents a two-stage methodology for SA in code-mixed sentences. In the Language Identification stage, a One vs One Multi-Class SVM classifier outperforms a rule-based approach with logistic regression, demonstrating a 6% increase in average F-score compared to Conditional Random Field (CRF) for language identification. For sentiment analysis, the study employs back transliteration of Indic languages through the *Google*® Translate Application Programming Interface and language-specific *SentiWordNets*. The proposed algorithm surpasses a machine-translated baseline, showing an 8% and 10% improvement in precision and recall, respectively.

Contributing to the understanding of mixed language sentiments, the study by [14] focused on addressing the challenges of SA in Sinhala-English code-mixed content. It emphasized the limitations of exclusively relying on Sinhala for SA due to the prevalent code-mixing in social media comments, which often include *Singlish* terms in Roman script. To overcome this challenge, the research developed a model using 500 code-mixed comments from *YouTube*®. These comments were manually labeled for sentiment as Positive, Negative, or Neutral. Additionally, a *Singlish* to Sinhala dictionary was created through transliteration, contributing to the understanding of mixed language sentiments. This study employed various feature extraction techniques and supervised machine learning methods, achieving a transliteration accuracy of 72%. Notably, the RF classifier demonstrated the highest accuracy at 75%, providing valuable insights for SA in Sinhala-English code-mixed content.

However, SA on Sinhala-English is a new approach, and there is no available dataset for *Singlish* language with sentiment annotations and thus created a sentiment dataset for *Singlish* with voluntary annotators using *YouTube*® comments. Also, shown the baseline results with machine learning algorithms using the Sinhala-English sentiment dataset..

III. Materials and Methods

A. *Corpus creation*

Comments from *YouTube*® videos on various fields are considered to create resources for Sinhala-English code-mixed text. As this study focuses on the sentiment polarity, videos that can be easily understood by the most of the users were chosen since the comments and the sentiments rely on the users. The collected data contains code-mixed comments of all three types; Inter-Sentential switching, Intra-Sentential switching, and Tag switching. Both Roman script and Sinhala script words used several types of mixing strategies.

Code-mixing has several types and it is not easy to match with monolingual data. In this work, create the new dataset as the first step in entering to Sinhala-English Code-mixed scenario and also present baseline classification results on the new dataset. This paper contributes providing the Sinhala-English code-mixed dataset annotated for SA and give results on some machine learning classification methods. Information from newspapers, novels, movies, and radio broadcasts to television shows, videos, and tweets can make a corpus. In natural language settings, corpus contains text and spoken data that can be used to train Artificial Intelligence and machine learning systems. The objective of the study is to create a corpus for Code-Mixed Sinhala-English language for SA related research. First, 140,582 comments were collected from various types of videos such as interviews, reality shows, etc. Those comments contained sentences that were either fully written in Sinhala or Code-mixed *Singlish* or fully written in English. And also, there were some comments in other languages such as Hindi-English. *YouTube*® comments are much noisy because users express their ideas naturally. In pre-processing all collected comments are cleaned to remove noises and extracted the *Singlish* data only.

All words which are not written in Roman scripts have been discarded and removed all emojis from the collected data as this study considered only *Singlish* words. Pre-process removed Hyper Text Markup Language (HTML) tags and Uniform Resource Locators (URLs) since those are not revealing any sentiments. Uppercase letters were also replaced using lowercase letters for simplicity of the dataset and easy use in the machine learning processes. Punctuations and numbers were also removed to focus on *Singlish* text data for sentiment classification. Finally, all comments were pre-processed to remove all unwanted white spaces and duplicate comments. In addition, every character repetition with more than two occurrences is replaced with two characters. Python programming language on the Google Collaboratory (*Google Colab*®) Notebook is used to implement these steps.

Data pre-processing involves the conversion of raw data into well-processed datasets, which can be used to perform data excavation analytics. After the pre-processing steps, comments were collected written in Roman scripts only.

B. Annotation setup

For annotation, partially used the Semantic-Role-based approach as described in [10] where each sentence is annotated by at least two annotators for the states of Positive (Pos), Negative (Neg), and Neutral (Neu). The study [10] addresses complexities in annotating sentiment beyond simple positive, negative, or neutral labels. It identifies challenging sentence types and proposes two annotation schemes. The Simple Sentiment Questionnaire provides concise instructions for annotating dominant sentiment, while the Semantic-Role Based Questionnaire offers a detailed approach considering the speaker's emotional state, primary target of opinion, and nuanced sentiment assessment. The former ensures cost-effectiveness, while the latter caters to diverse application needs. Practitioners are encouraged to adapt these questionnaires for specific applications.

This study considered the mixed feeling state as neutral since it does not give a clear idea about any sentiment. All comments were processed as anonymous to keep the privacy of the people commented. Annotators were informed about the annotation schema and the purpose of the data. As pre-processed data have comments in Roman script, they were asked to annotate the comment not entirely in English or not giving any meaning.

The comments with less than three words and longer comments were also skipped to avoid difficulties of the prediction of the sentiment during the analysis. All the data are divided into spreadsheets with a maximum of 500 comments and ask annotators to choose the files as they prefer. The annotation process is performed in three steps. Firstly, annotate all comments as followed in the selected schema with the status Pos, Neg and Neu. After the first annotation step, filtered out the comments entirely written in English and the comments which do not give a clear idea about the sentiment. After each comment was annotated by two persons, the data were collected if both of them agreed. In the case of a conflict, the third person annotated the comment. If all the three of them have not indicated the same annotation, then another annotation is considered from the fourth annotator and selected the most commented annotation.

C. Annotators

A selected group of annotators was requested to contribute annotating the preprocessed comments. In this work chooses different annotators on the basis of gender, educational level and medium of higher education as listed in TABLE I. All of them were diverse Sinhala native speakers in different professional fields. All the annotators have enough conversational level of understanding Sinhala and English language fluently. Each annotator was informed clearly with regard to the annotation setup to align with the schema.

D. Corpus statistics

For the Sinhala-English code-mixed dataset, corpus statistics are shown in TABLE II. As mentioned in section B, all data have been categorized into three groups: Positive, Negative, and Neutral. The distribution of the final dataset is shown in TABLE III. The dataset may be biased into the positive state since most of the users choose to leave a comment when they like *YouTube*® videos and they have not given any comment if they do not like probably. The number of comments annotated in the dataset for each of the sentiments are tabulated in TABLE III. The created dataset was split into two, one for training with 6,265 *Singlish* comments and another for testing with 1,567 tweets.

TABLE I: ANNOTATORS

Factor	**Categories**	**Number of Annotators**
Gender	Male	02
	Female	07
Educational Level	Non-degree	02
	Undergraduate	04
	Graduate	03
Medium of the highest education	English	06
	Sinhala	03
Total		**09**

TABLE II: CORPUS STATISTICS OF SINHALA-ENGLISH DATA

Statistic Factor of Corpus	**Number of Sinhala-English Comments**
Number of tokens	74,725
Number of Unique tokens	17,111
Number of Comments	7,832
Average sentence length	10

TABLE III: DATA DISTRIBUTION OF CREATED SINHALA-ENGLISH DATASET

Sentiment Category	**Number of Sinhala-English Comments**
Negative	1,853
Positive	4,263
Neutral	1,716
Total	7,832

E. Benchmark systems

As machine learning algorithms have been evolving, the general algorithms have become smarter to automatically learn by the machine from the given data. This paper presents insight of the newly created *Singlish* code-mixed language dataset using machine learning algorithms. The dataset was divided into 80% for the training phase and 20% for the testing case as most commonly practiced in machine learning algorithms. The dataset was evaluated using five machine learning models: SVM, LR, DT, RF, and KNN. This thorough approach not only showcases the diversity of feature extraction methods but also emphasizes the careful configuration of each machine learning model, ensuring a comprehensive evaluation tailored to the intricacies of the Sinhala-English code-mixed language dataset with sentiment annotation.

In this work, applied the pipeline with several feature extractions. The study chooses the TF IDF as the input feature with the value of up to two grams for better results. It is about how often a word appears in a document, compared to how many words are there. LR model implemented with a higher number of maximum features while other methods (SVM, DT, RF) with 2500 maximum features. KNN model applied without specifying the maximum number of features with 1, 2, 3, and 5 neighbours. The library *Scikit-learn* (*Sklearn*) was used for evaluation on *Google Collaboratory*®.

IV. Results and Discussion

In this study, evaluated the created Sinhala-English Code-Mixed Dataset based on the precision, recall, and F-Score of selected baseline models. The experimental results of the sentiment classification using five models are shown in TABLE IV. Precision discusses the total predictive positives and Recall is a metric that illustrates how many of the actual positives, captured through labeling them as positive. F1-score considers a better measurement to seek a balance between the precision and recall with an uneven class distribution.

The macro average calculates the metric with class independence and then takes the unweighted mean when an imbalance of data sources exists. The weighted average calculates the metrics for each of the classes independently. The created test dataset contains 856 positive comments, 364 negative comments, and 347 neutral comments.

As shown in TABLE IV, all the used classification algorithms achieved considerable accuracy in creating the *Singlish* dataset. LR shows the higher macro average score for precision, recall, and F1 score as 0.69, 0.64, and 0.65 respectively. The extracted values of macro average, weighted average, and accuracy for the F1 score is shown in TABLE V to make comparison of the five models.

These five algorithms show that the F1 Score, Recall, and Precision are higher for the Positive class because the dataset could be biased to Positive polarity. In this study collected 54% of comments belong to Positive class while the other classes Negative and Neutral has 24% and 22% respectively.

To showcase the dataset's potential insights, our study employs five distinct machine learning algorithms, achieving noteworthy accuracy in sentiment classification. This contribution not only addresses the scarcity of resources in the Sinhala-English code-mixed language but also opens avenues for future research and development in the field of sentiment analysis within the multilingual contexts. While these models yielded promising results, the findings presented here provide a foundational understanding and set the stage for further exploration and refinement in Sinhala-English code-mixed sentiment analysis. As the field evolves, incorporating more sophisticated techniques and diverse datasets will undoubtedly contribute to the ongoing progress in this burgeoning area of research.

Table IV: The experimental results of the sentiment classification

		Precision	**Recall**	**F1-score**
SVM	Neg	0.69	0.61	0.65
	Neu	0.6	0.34	0.43
	Pos	0.75	0.92	0.82
	accuracy			0.71
	macro avg	0.68	0.62	0.63
	weighted avg	0.7	0.71	0.69
LR	Neg	0.67	0.64	0.65
	Neu	0.63	0.37	0.46
	Pos	0.76	0.91	0.83
	accuracy			0.72
	macro avg	0.69	0.64	0.65
	weighted avg	0.71	0.72	0.71
DT	Neg	0.55	0.54	0.54
	Neu	0.46	0.42	0.44
	Pos	0.75	0.78	0.77
	accuracy			0.65
	macro avg	0.59	0.58	0.58
	weighted avg	0.64	0.65	0.64
RF	Neg	0.61	0.57	0.59
	Neu	0.54	0.35	0.42
	Pos	0.75	0.88	0.81
	accuracy			0.69
	macro avg	0.63	0.6	0.61
	weighted avg	0.67	0.69	0.67
KNN	Neg	0.59	0.52	0.55
	Neu	0.45	0.37	0.4
	Pos	0.73	0.83	0.78
	accuracy			0.65
	macro avg	0.59	0.57	0.58
	weighted avg	0.64	0.65	0.64

Table V: Comparison of the results of five models

	SVM	**LR**	**DT**	**RF**	**KNN**
Accuracy	0.71	**0.72**	0.63	0.69	0.65
Macro average	0.63	**0.65**	0.56	0.61	0.58
Weighted average	0.69	**0.71**	0.63	0.67	0.64

V. Conclusion

The study presents the first Sinhala-English Code-Mixed dataset with annotation for sentiment polarity using *YouTube*® comments. While the dataset holds immense value, the inherent imbalance with three labels raises concerns about potential biased learning in models. To mitigate this, we advocate for future research to delve into various sampling techniques, including undersampling, oversampling, or a combination of both. This exploration is crucial to rectify the imbalance and elevate the overall performance of sentiment analysis models. Recognizing the probable bias stemming from users' inclination to predominantly express positive sentiments, we emphasize the importance of transparently addressing these tendencies. This acknowledgment enhances the credibility of our findings.

Additionally, our in-depth analysis advocates for the incorporation of advanced sampling techniques to further fortify model robustness, ensuring equitable representation across classes in the Sinhala-English Code-Mixed Dataset. The evaluation of machine learning models on the newly created Sinhala-English Code-Mixed dataset underscores the potential of these models. Particularly, the Logistic Regression model exhibits superior performance across precision, recall, and F1-score metrics. This analysis provides valuable insights for future research and applications in the dynamic landscape of Code-Mixed sentiment analysis. Furthermore, the study underscores the need for ongoing vigilance in monitoring and adapting models to evolving language trends and user behaviors on platforms like *YouTube*®. As online communication patterns shift, sentiment analysis models should remain dynamic and responsive to maintain relevance and accuracy.

In conclusion, while this study represents a significant milestone in sentiment analysis for Sinhala-English Code-Mixed data, the journey towards effective modeling is ongoing. Continued interdisciplinary collaboration, innovation in sampling techniques, and a commitment to cultural sensitivity collectively contribute to refining sentiment analysis models in this distinctive linguistic landscape. The presented dataset is available at *https://github.com/1516CS/Sentiment-dataset-Singlish.git*, for further exploration and collaboration, enabling researchers to address new challenges in the evolving field of Code-Mixed research within the Sinhala-English language scenario.

REFERENCES

[1] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, and J. P. McCrae, “A sentiment analysis dataset for code-mixed Malayalam-English,” arXiv preprint arXiv:2006.00210, 2020.

[2] N. Medagoda, S. Shanmuganathan, and J. Whalley, “Sentiment lexicon construction using SentiWordNet 3.0,” IEEE 11th International Conference on Natural Computation (ICNC) 2015, pp. 802-807, August 2015.

[3] P. D. T. Chathuranga, S. A. S. Lorensuhewa, and M. A. L. Kalyani, “Sinhala sentiment analysis using corpus-based sentiment lexicon,” IEEE 19th international conference on advances in ICT for emerging regions (ICTer) 2019, vol. 250, pp. 1-7, September 2019.

[4] A. Pravalika, V. Oza, N. P. Meghana, and S. S. Kamath, “Domain-specific sentiment analysis approaches for code-mixed social network data,” IEEE 8th international conference on computing, communication, and networking technologies (ICCCNT) 2017, pp. 1-6, July 2017.

[5] G. I. Ahmad, J. Singla, and N. Nikita, “Review on sentiment analysis of Indian languages with a special focus on code mixed Indian languages,” IEEE International Conference on Automation, Computational and Technology Management (ICACTM), pp. 352-356, April 2019.

[6] M. A. Ansari, and S. Govilkar, “Sentiment analysis of mixed code for the transliterated Hindi and Marathi texts,” International Journal on Natural Language Computing (IJNLC), vol. 7, 2018.

[7] S. Mukherjee, “Deep learning technique for sentiment analysis of Hindi-English code-mixed text using the late fusion of character and word features,” IEEE 16th India Council International Conference (INDICON) 2019, pp. 1-4, December 2019.

[8] S. Lee, and Z. Wang, “Emotion in Code-switching Texts: Corpus Construction and Analysis,” Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, pp. 91-99. DOI: 10.18653/v1/W15-3116, 2015.

[9] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae, “Corpus creation for sentiment analysis in code-mixed Tamil-English text,” arXiv preprint arXiv:2006.*00206*, 2020.

[10] S. Mohammad, “A practical guide to sentiment annotation: Challenges and solutions,” Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment, and social media analysis, pp. 174-179, June 2016.

[11] R. Bhargava, Y. Sharma, and S. Sharma, “Sentiment analysis for mixed script indic sentences,” IEEE International conference on advances in computing, communications and informatics (ICACCI) 2016, pp. 524-529, September 2016.

[12] A. Konate, and R. Du, “Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques,” Wuhan University Journal of Natural Sciences, vol. 23(3), pp. 237-243, 2018.

[13] S. Mandal, S. K. Mahata, and D. Das, “Preparing Bengali-English code-mixed corpus for sentiment analysis of Indian languages,” arXiv preprint arXiv:1803.04000, 2018.

[14] P. M. I. U. Aththanayaka, and H. M. M. Naleer, “Sentimental analysis of comments in social media in sinhala-english code-mixed language using supervised learning techniques,” Ninth Annual Science Research Sessions (ASRS) 2020, pp. 24, 2020.