Grenze International Journal of Engineering and Technology, Jan Issue

A Proposed Method to Translate Code Blended Web based Text in English Language

Sudeshna Sani[1], Dipra Mitra[2], Kumar Gaurav[3], Kanika Thakur[4] Srihari Babu Gole[5] and Souradeep Sarkar[6]
[1]Woxsen University/ Department of Business Analytics, SOB, Telangana, India
Email: sudeshnasani@gmail.com
[2-4]Amity University/ Department CSE, Ranchi, India
Email: Mitra.Dipra@gmail.com, {kgaurav, kthakur }@rnc.amity.edu
[5]Anurag University/ Department of Information Technology, School of Engineering, Hyderabad, Telangana, India
Email: sriharibabug@gmail.com
[6]Brainware University/ Dept.of CSS, Kolkata, India
Email: sos.cs@brainwareuniversity.ac.in

***Abstract*— Code-blended online texts are increasingly prevalent due to multilingual influences. This phenomenon is commonly observed in user-generated content on social platforms, especially from multilingual users. Such content, including articles and information, possesses informal language features like non-standard abbreviations, contracted transliterations, and casual grammar. To effectively handle and analyze code-blended data for various Natural Language Processing tasks, understanding this phenomenon is crucial. As the need for translating code-blended content into standard language grows, our study focuses on short utterances gathered from Facebook, Twitter, and WhatsApp. The dataset, comprised of English and Marathi language pair from Indian web-based text, aims to provide a resource for translating code-blended content into plain English. We have tested the dataset on various machine learning classifiers, achieving good accuracy in language identification, and thereafter translation was performed through LSTM network. We scored 91% prediction accuracy while translating code mixed sentences into normal English language sentences.**

***Index Terms*— Code-blending, multilingual interference, transliteration, SMO classifier, LSTM.**

I. INTRODUCTION

Correspondence is significant for our day-to-day existence. To speak with one another, we want language as an instrument of correspondence. While concentrating on language, we are emphasizing human essence which is the unmistakable characteristic of the brain that fluctuates from one man to another [1]. In recent decades, advanced correspondence mediums like email, Facebook, Twitter, WhatsApp, SMS, and so forth have enabled individuals to have discussions in a considerably more casual way than previous days' communication mediums. Web-based life content contrasts in other traditional content from numerous points of view. The online networking correspondence, a multilingual speaker frequently utilizes more than one language. This very casual nature of discursion has offered to ascend to another type of cross-breed language, called code blended language which does not have an officially characterized structure.

Being influenced by circumstances and people meeting around, by and large, people frequently utilize more than one language. The peculiarity of individuals talking and seeing more than one language is called bilingualism or

Grenze ID: 01.GIJET.10.1.93_1

multilingualism [2]. Code blending happens when individuals blend two dialects (or more) in their discourse act or talk with practically no power to blend code. Code blending happens when people blend (at least two) vernaculars in their correspondence with no expectation of blending codes. It has turned into a typical peculiarity in a bilingual and multilingual society. Code-Blending insinuates the inclusion of phonetic units, for instance, articulations, words or morphemes of one language into an explanation of another tongue [3]. As an issue of significance, we need to appreciate the need of language exchanging. Customized distinctive evidence of the code exchanging is huge as it understands the repeat of code trade or code blending. Code blending moreover consolidates short articulations of certified words found in electronic life compositions.

With the demand for translating code-blended content into conventional language rising, our research zeroes in on concise utterances sourced from Facebook, Twitter, and WhatsApp. The dataset, encompassing English and Marathi language pairs from Indian web-based text, seeks to offer a solution for converting code-blended content into clear English. We assessed the dataset using diverse machine learning classifiers, attaining favorable language identification accuracy. Subsequently, through LSTM [4] network, we achieved a remarkable 91% accuracy in transforming code-mixed sentences into standard English equivalents.

II. Related Works

Bali et al., *2014* gathering carried out assessment of data from Facebook posts made by Hindi-English bilingual clients. In this paper, they present an examination of information from Facebook produced by EnHin bilingual clients. Their examination uncovers which a lot of this information had Code Blending as English lattice as well as Hindi network[3].

Barman et al., 2014 portrayed their work underway on the issue of automatic language ID for the different tongues of virtual entertainment. Author gives another dataset which contains Facebook posts and remarks show code blending between Bengali, English and Hindi [5].

Gupta et al., 2016 proposed to develop a structure for language-recognizing evidence and word standardization for Hindi-English code-blended internet organizing content. They have given clarification rules to the system, directly following researching the many-sided thought of the data assortment used [6].

Ghosh et al., 2017 has played out an AI based feeling grouping of Facebook posts. The extremity of each post has been named positive, negative and unbiased. They have made their very own dataset. They performed opinion examination of code-blended online text. The best result was obtained using a blend of word-based and semantic features with an accuracy of 68.5 [7].

Bohra et al., 2018 accumulated the corpus of Hindi-English code blended content, containing different tweets and their connected remarks. Moreover they introduced the coordinated system used for distinguishing proof of Disdain Discourse in the codeblended content. They had separated the issue of disdain discourse recognition in code blended messages and introduced a Hindi-English code-blended dataset containing tweets posted web-based on Twitter [1].

Asubiaro et al., 2018 have fostered a methodology for naturally recognizing and segregating the language of words in a bilingual text. The outcomes indicate that the proposed methodology is language-free and it requires a little corpus to accomplish phenomenal accuracy and review [8].

Sani et al. (2023) developed a sentiment analysis model for multilingual tweets using NLP techniques. They utilized different classifiers (SNN, CNN, and LSTM) to identify positive and negative comments, aiming to enhance classification accuracy. Their experiments showed that the LSTM Neural Network and CNN+LSTM combination achieved the highest accuracy. The sentiment analysis dataset had an 84.1% accuracy in classifying positive and negative sentences, with a reasonable average time complexity [4].

Abdullah et al.(2022) conducted research on emotion detection in code-mixed texts, involving both Roman Urdu (RU) and English (EN), which were gathered from social media platforms. They created the RU-EN-Emotion corpus using a meticulous and structured procedure. The corpus comprises 20,000 RU-EN code-mixed sentences extracted from a collection of 400,000 sentences acquired from three distinct sources. The selected sentences were subjected to manual annotation at two levels [9].

III. Proposed Work

The code-blended data was manually selected from various web-based texts and messaging sources. The following figure shows the proposed methodology. And the rest of the paper explains every step.

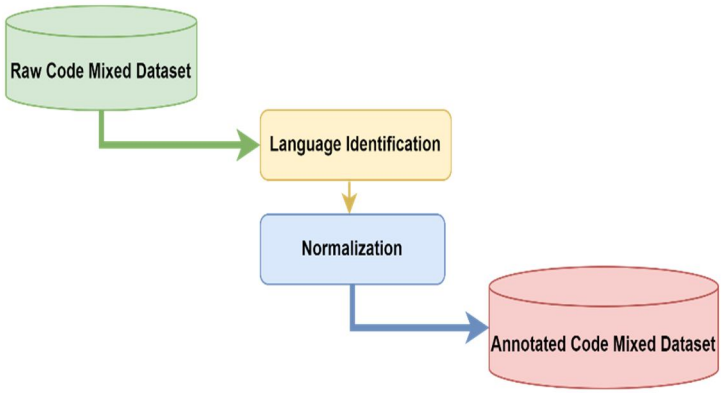

Figure 1. Schematic Diagram of Translation of code blended Dataset

A. Raw code blended Dataset

Data Source: In this paper, information sources were derived from the discussions among participants. The researchers obtained additional information from these sources to collect the necessary data. For example, online media platforms like Facebook, Twitter, WhatsApp, etc., can serve as information sources for gathering data.
Data Collection: The information is gathered by utilizing the narrative method. In this cycle, we utilize a smart technique to get legitimate and genuine information. We have chosen the WhatsApp (WA) chat, Facebook Messenger, and Twitter Tweets to find out the code blending and collected some relevant material of code blended texts from there.

Figure 2. Marathi-English code blended Text on Facebook

Data Analysis : Data analysis [10] is a procedure for getting crude information and changing it into data which will be helpful for basic leadership by clients. It includes working with information, sorting and breaking it into sensible units, and incorporating and finding what is significant.

Figure 3. Phases in Dataset Creation

Extraction of code blended Text: code blended texts are presented in a raw collection of data that needs to be extracted for further processing. A variety of code-blended text is present in social media conversations. We extracted code-blended data from social media contexts such as Facebook, WhatsApp, and Twitter for further processing.

Code blended Dataset: In this step, after extraction, we get the relevant code blended dataset which needs to be organized properly. The code-blended data was manually selected from various social media and messaging sources. Here we used English code blended sentences and dataset is annotated with language tag as EN for English. Dataset example the phrase "HAVE A GREAT DAY!!" can be written in multiple ways, like Have a gr8 day. Hv a grt day. Have a great day.

B. Language Identification

Distinguishing language in online person-to-person communication posts has specific hardships connected with it. Spelling blunders, phonetic composing, utilization of transcribed letters in order, and shortenings make this fascinating. When further combined with the problem of transliterated code-blending, it becomes even harder to disambiguate words of one language from another. Language identification involves the following steps shown in Fig-4.

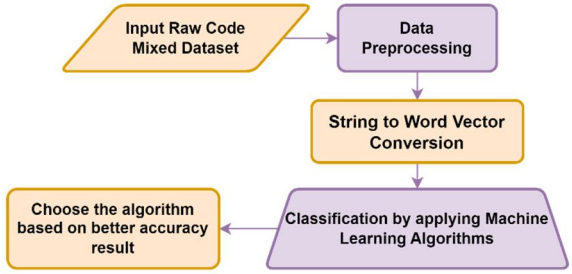

Figure 4. Steps involved in Language Identification

Data Preprocessing: A system is used to change over the unrefined information into an ideal educational record. specific word information is consistently divided into unambiguous practices or inclinations and is likely going to contain various mistakes. Information preprocessing is a shown methodology for settling such issues. We want information preprocessing because of understanding reasons. The content contains advanced punctuation and lowercase [11]. There are outstanding characters in the Marathi. There are topics discussed in English with various interpretations in Marathi. Records are referenced by sentence length with bigger sentences close to the finish of the report.

Clean Text: At first, we ought to stack the data with the end goal which defends the Unicode Marathi characters. Each line holds single sets of articulations, first English and thereafter Marathi-English, disengaged by a tab character. We ought to part the stacked substance by line and subsequently by express. Content requirements to clean each sentence [12]. We eliminated all accentuation characters and normalized all Unicode characters to ASCII. Standardize the case to lowercase. Eliminate any extra tokens which are not alphabetic. We play out these methods on every articulation for each pair in the stacked dataset. Following performing cleaned movement, we can save the overview of articulation sets to a record prepared for use. We use the pickle Programming interface to save the once-over of clean happy to record.

```
Saved: english-marathi.pkl
[Ok Go.] => [Ok जा.]
[Run fast.] => [पळ fast.]
[Run fast.] => [धाव fast.]
[Run fast.] => [पळा fast.]
[Run fast] => [धावा fast.]
[Who?] => [कोण?]
[Wow this is nice.] => [वाह this is nice.]
[Fire Shitt.] => [आग Shitt.]
[Fire Shitt.] => [फायर Shitt.]
[Help please.] => [वाचवा please.]
[Help please.] => [वाचव please.]
[Jump Now.] => [उडी मार Now.]
[Jump Now.] => [उडी मारा Now.]
[Jump Now.] => [उडी मार Now.]
[Jump Now.] => [उडी मारा Now.]
[Stop please.] => [थांबा please.]
```

Figure 5. Clean Dataset

```
@attribute Back numeric
@attribute Bae numeric
@attribute Battery numeric
@attribute Be numeric
@attribute Before numeric
@attribute Best numeric
@attribute Boyfriend numeric
@attribute Bring numeric
@attribute Brother numeric
@attribute Bst numeric
@attribute Bye numeric
@attribute C numeric
@attribute C&P numeric
@attribute CU numeric
@attribute CYA numeric
@attribute Calculate numeric
@attribute Call numeric
```

Figure 6. String To Word Vector Conversion of Text

String to Word Vector Conversion: In the preprocessing step, we need to transform the text strings into a term vector to enable learning. This is done by applying the String to Word Vector filter which is the most remarkable text mining function which is an unsupervised filter attribute. Tokenization is the way to separate content into

significant pieces. These pieces are called tokens. For instance, we can isolate a piece of content into words, or we can separate it into sentences.
Classification: For language identification and classification we have employed Sequential Minimal Optimization (SMO), Logistic Regression and Multiclass classifier algorithms [13]. We needed to check the language identification accuracy of our dataset by testing our dataset. Table 1 shows the accuracy results for the dataset tested on machine learning classifiers. With SMO we got the best accuracy for our English-Marathi Dataset.

TABLE1. MODEL ACCURACIES FOR LANGUAGE IDENTIFICATION

Model	Accuracy
SMO	70.1525 %
Logistic Regression	69.3682 %
Multi-class Classifier	68.061 %

C. Translation

Translation is a process that converts a list of words into a more consistent sequence, which is useful for text preparation in subsequent training. When training a translation model, machine translation systems are considered, and pre-processing is required to adapt them to our code-blended dataset. In the language identification task, it was necessary to translate nonstandard tokens in the text to standard tokens. To address this, a translation module was implemented to perform language-specific adjustments and provide the correct spelling for each word. With the proliferation of SMS and online social media content, text normalization systems have gained attention, focusing on transforming these tokens into standard dictionary words.
Split dataset into train and test data: The spotless data contains somewhat more than 1200 sets. The complex idea of the model augmentations with the number of examples, length of articulations, and size of the jargon. Notwithstanding the way in which we have a dataset for showing interpretation, we changed the issue conceivably to diminish the size of the model required, and thus the arranging time expected to fit the model. We have revised the issue by reducing the dataset to the underlying 1000 models in the record [14]. Further, we have taken 700 of those as models for preparation and the remaining 300 advisers for testing the fit model.
Train Neural Translation Model: This solidifies both stacking and setting up the ideal substance data figured out for showing up and portraying and setting up the model on the set-up data. We will utilize the "both" or blend of the training and testing datasets to depict the most phenomenal length and vocab of the issue. This is for straightforwardness. On the other hand, we could portray these properties from the status dataset alone and gather models in the test set which are exceptionally wide or have words that are out of the language.
We utilized Keras Tokenizer class for the numerical representation of entire numbers in English and Marathi phrases. Separate tokenizers were used for each language. The combined dataset determined vocabulary sizes and maximum sequence lengths. To prepare the data, input and output sequences were encoded and padded. Word embedding was applied to input sequences, while output sequences were one-hot encoded to capture word probabilities. We implemented an encoder-decoder LSTM model [15]. The encoder encoded input sequences, and the decoder decoded them word by word. Model architecture and parameters such as vocabulary sizes, sequence lengths, and memory units were defined. The model was trained using the Adam optimization algorithm for 75 epochs, treating it as a multiclass classification problem. A batch size of 64 models was used.

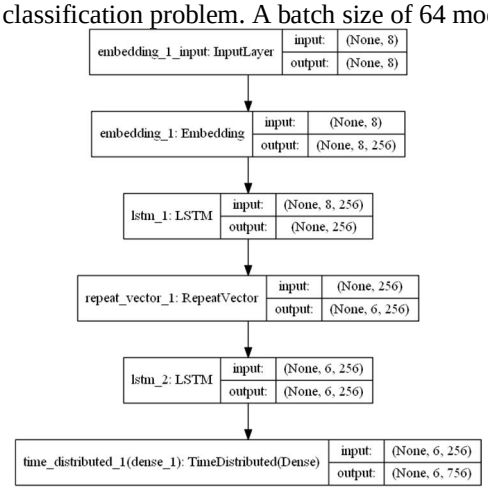

Figure 7. Plot of Model Graph for Neural Translation Model

```
Train

src=[आम्ही responsible नाही.], target=[We're not responsible.], predicted=[we're not responsible]
src=[टॉमला eggs आवडत नाहीत.], target=[Tom doesn't like eggs.], predicted=[tom doesn't like eggs]
src=[आम्हाला ते कधीतरी दिसतात.], target=[We sometimes see them.], predicted=[we sometimes see them]
src=[आम्हाला आणखीन one year आहे.], target=[We have one more year.], predicted=[we have one more year]
src=[ते बोलले rudely.], target=[He spoke rudely.], predicted=[he spoke rudely]
src=[टॉम २०१३ मध्ये graduate झाला.], target=[Tom graduated in 2013.], predicted=[tom graduated in 2013]
src=[टॉमला math आवडत नाही.], target=[Tom doesn't like math.], predicted=[tom doesn't like beef]
src=[टॉम मेरीसाठी work करतो का?], target=[Does Tom work for Mary?], predicted=[tom tom tom a mary]
src=[आम्ही जिंकलो finally.], target=[Finally we won.], predicted=[we we won]
src=[लवकर.], target=[Hurry.], predicted=[hurry]
src=[टॉम जिंकू शकत नाही We know], target=[We know Tom can't win.], predicted=[we know tom can't win]
src=[तो धावला Yes.], target=[Yes He ran.], predicted=[yes he ran]
src=[तो bus ने आला.], target=[He came by bus.], predicted=[he came by bus]
```

Figure 8. Translated Result for Train Dataset

```
Test

src=[ती मेली, Oh no.], target=[She died, Oh no.], predicted=[she died oh no]
src=[टॉम october मध्ये वारला.], target=[Tom died in October.], predicted=[tom died in october]
src=[टॉमचा dog खूप भुंकतो.], target=[Tom's dog barks a lot.], predicted=[tom's dog barks a lot]
src=[टॉम निघून गेला quickly.], target=[Tom left quickly.], predicted=[tom left quickly]
src=[टॉमला सावध कर now.], target=[Warn Tom now.], predicted=[warn tom now]
src=[टॉमला माहीत होतं already.], target=[Tom already knew.], predicted=[tom already knew]
src=[तुम्ही guitar वाजवता का?], target=[Do you play the guitar?], predicted=[do you play the guitar]
src=[तिला help ची गरज आहे.], target=[She is in need of help.], predicted=[she is is to of help]
src=[आता कोणीही आपली help करू शकत नाही.], target=[No one can help us now.], predicted=[tell one can help us now]
src=[टॉमला मी tomorrow call करेन.], target=[I'll call Tom tomorrow.], predicted=[i'll call tom tomorrow]
src=[टॉम bananas विकत घेत आहे.], target=[Tom is buying bananas.], predicted=[tom is buying bananas]
src=[मी mountains मध्ये होते.], target=[I was in the mountains.], predicted=[i was in the mountains]
src=[उठ fast.], target=[Stand up fast.], predicted=[stand up fast]
```

Figure 9. Translated Result for Test Dataset

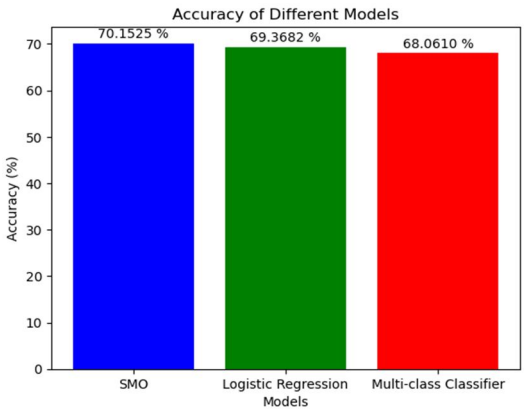

Figure 10. Language Identification with SMO

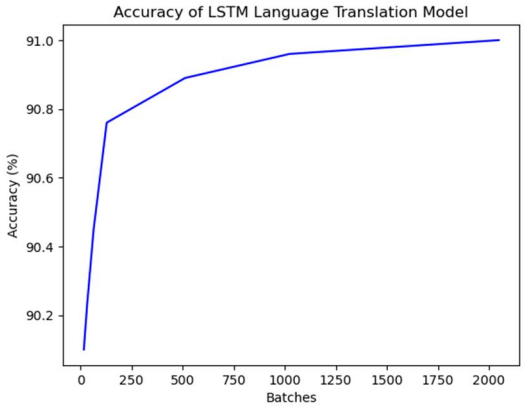

Figure 11. Accuracy Graph for LSTM model

IV Conclusion and Future Works

We've focused on automating language identification and text standardization through the integration of Indian language code-mixing from virtual communication in our study. This language identification task is quite intricate, requiring word-level analysis, as messages and sentences can consist of words in multiple languages. We've identified several types of code-mixing instances in the data, including word inclusion, phrase inclusion, state inclusion, and clause inclusion. Our dataset, created by analyzing raw data from diverse virtual communication sources, encompasses various forms of code-mixing. The paper underscores the challenges posed by code-blended text in language processing. The current dataset includes approximately 2000 instances of Marathi and English code-mixing words, converted into English by initially identifying the language of the dataset. Table 1 displays the results of testing the dataset on various machine learning classifiers, achieving good accuracy in language identification, followed by standardization for designated language text through LSTM network. We scored 91% accuracy while translating code mixed sentences into normal English language sentences.

REFERENCES

[1] Bohra, Aditya Vijay, Deepanshu Singh, Vinay Akhtar, Syed Sarfaraz Shrivastava, Manish. . A Dataset of Hindi-English Code- Mixed SocialMedia Text for Hate Speech Detection. 36-41. 10.18653/v1/W18-1105. 2018

[2] Chakraborty, IV Koyel, Sudeshna Sani, and Rajib Bag. "A Study On Sentiment Polarity Identification Of Indian Multilingual Tweets Through Different Neural Network Mod-els." J. Mech. Cont. Math. Sci 15.1,2020 p.p: 108-117

[3] Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas, "I am borrowing ya blending? an analysis of english-hindi code blending in facebook", In proceedings of the First Workshop on Computational Approaches to Code Switching, October, 2014.

[4] Sani, Sudeshna Mitra.Dipra , Mondal, Soumen. (2023). Detection of Polarity in the Native-Language Comments of Social Media Networks.10.1201/9781003328414-22.

[5] Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster, "Code blending: A challenge for language identification in the language of social media", In proceedings of the first workshop on computational approaches to code switching, October 2014.

[6] Gupta, Sakshi Bansal, Piyush Mamidi, Radhika.. Resource Creation for Hindi-English Code Mixed Social Media Text,2016.

[7] Ghosh, Souvick Ghosh, Satanu Das, Dipankar. Sentiment Identification in Code-Mixed Social Media Text.ToluwaseAsubiaro, 2017

[8] Toluwase Asubiaro, Robert E. Mercer, Isola Ajiferuke, "A Word-Level Language Identification Strategy for Resource-Scarce Languages",In conference of the Association for Information Science and Technology, Vancouver, BC, Canada, November 2018

[9] Ilyas, Abdullah Shahzad, Khurram Malik, Muhammad Emotion De-tection in Code-Mixed Roman Urdu - English Text. ACM Transactions on Asian and Low-Resource Language Information Processing. 22. 10.1145/3552515,2022.

[10] D. Mitra, N. Sharma, M. Rashid and R. Singh, "Classification Rules based Breast Cancer Detection using Machine Learning Approach,"2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 1274-1278, doi:10.1109/IC3I56241.2022.10072832.

[11] Mitra, D., Gupta, S. (2022, April). Plant Disease Identification and its Solution using Machine Learning. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 152-157). IEEE

[12] Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhat- tacharyya, "Multitasking of sentiment detection and emotion recognition in codemixed hinglish data. Knowledge-Based Systems", 2023

[13] Sani, S., Bera, A., Mitra, D., Das, K. M. (2022). COVID-19 Detection Using Chest X-Ray Images Based on Deep Learning. International Journal of Software Science and Computational Intelligence (IJSSCI), 14(1), 1-12

[14] Mitra, D., Gupta, S., & Goyal, A. (2022, April). Security system using open cv based facial recognition. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 371-374). IEEE.

[15] Zilpe, Shradha & Sani, Sudeshna & Mitra, Mr.Dipra & Ghosh, Abhinandan & Kumar, Pranav & Singh, Amarnath. (2023). A review of object–based detection using convolutional neural networks. 10.1201/9781032684994-70.