International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7840
***Article Received**: 26 April 2023 **Revised**: 20 June 2023 **Accepted**: 01 July 2023*

Survey on Hinglish to English Translation and Classification Techniques

Nicole D'Souza[1], Devarsh Patel[2], Jigyashu Saravta[3], Dr. Ashwini Rao[4]
[1]Student, MPSTME, IT Dept, NMIMS University
[2]Student, MPSTME, IT Dept, NMIMS University
[3]Student, MPSTME, IT Dept, NMIMS University
[4]Assistant Professor, MPSTME, IT Dept, NMIMS University
Mumbai (M.H.), India
ashwini.rao@nmims.edu

Abstract—Code-mixing is the process of using many languages in one sentence and has a widespread occurrence in multilingual communities. It is particularly prevalent in texts on social media. Due to the widespread usage of social networking sites, a substantial amount of unstructured text is produced. Hinglish, i.e. code-mixed Hindi and English, is a frequent occurrence in everyday language use in India. Hence, a translation process is required to help monolingual users and to aid in the comprehension of language processing models. In this paper, we study the effective techniques for classification and translation tasks and also find gaps and challenges in the current research domain. After comparing a few existing methodologies for machine translation, a framework which showed an improvement in task of translation over the previous methods is proposed.

Keywords- Code-mixing; hinglish; Natural language processing; Transliteration; Machine translation.

I. INTRODUCTION

India is a country with a variety of languages, and this, combined with a long history of international ties, has led to a bilingual society where most people prefer to speak in code-mixed languages during casual conversations. Code-mixing is an event when a speaker transitions between two or more languages in a single written or spoken sentence [1]. It does not have a set structure and often differs from person to person. The most widely used of these code-mixed languages in India is Hinglish, which is a mixture of Hindi and English and is mostly used in conversations on social media.

As a consequence of growing user participation on social networking sites, there is a rise in data and curiosity in studying and developing mechanisms that allow code-mixing of several Indian languages that are resource-constrained [1]. We have seen that individuals don't worry too much about communicating monolingually while speaking on social networking sites like Facebook, WhatsApp, Snapchat, and Twitter; instead, they end up using a mixture of languages [2]. Hinglish doesn't follow fixed standards for spelling and grammar users simply employ the phonetics of a word to come up with spelling. For instance, "Yes" in Hindi translates to "हाँ" which in Hinglish can be written as "ha", "haa", "haan", etc, based on a variety of reasons like regional pronunciations or dialectal conventions. Moreover, unlike monolingual languages, there is no formal data in the form of news articles or books. This results in researchers having to resort to using unstructured data from social media comments and posts [3].

In order to improve communication and information sharing with other nations, states, and central governments, there is a desperate and enormous need for improved machine translation systems. It is important to develop an efficient machine translation system so as to understand the correct meaning and sentiment of a given text. This translation is useful for a company to understand its customer review, it is useful for a user to understand comments on social media, etc.

Machine translation is a key area of research in the field of Natural language processing. It is a computerized and automated concept in charge of converting text from one language (referred to as the source language) to another (called target language) [4]. Machine translation of languages is a difficult process due to vocabulary shortage, context misinterpretation, syntax, semantics, grammatical errors, bias, pragmatics, phonetics, morphology, and other difficulties. As a result, handling code-mixed data becomes considerably more challenging since the composite structure creates several additional problems. Systems for machine translation deal with issues related to linguistic variation and ambiguity under the aegis of natural language processing.

The research community often uses the words code-mixing and code-switching interchangeably even though they are comparable yet formally distinct from one another. Let us first understand the difference between the two. (a) refers to code-mixing and (b) is an example of code-switching.

a) Rohan ek acha student hai.

Here, a mixture of Hindi and English words are used in the same sentence. This is code-mixing, also termed as Intra Sentential switching.

b) Rohan is a good student, aur voh khel khud mei bhi acha hai.

Above, we moved from English to Hindi while transitioning from the first sentence to the second one. This is code-switching which is also known as Inter-Sentential switching.

Below, is an instance of a Hinglish sentence and its translation.

- Sentence: My mother always told me ki mei bada aadmi banuga if I study.
- Gloss: [My mother always told me] that I big man become [if I study].
- Translation: My mother always told me that, I will become a big man if I study.

A. *Terminologies*

- ❖ BLEU (Bilingual Evaluation Understudy) Score: BLEU score is one of the most frequently used NLP metrics. It is based on the idea that the more similar the predicted content is to the target text that was produced by humans, the more accurate it will be. This score ranges from 0 to 1. A score around 0.6 and 0.7 is considered to be really good. Different individuals would likely come up with several different translations for a certain sentence and would rarely end up with a perfect fit. This makes a score closer to 1 irrational in actuality and should act as a red flag that your model is overfitting.
- ❖ N-Gram: An "n-gram" is not unique to NLP or BLEU Score; it is a commonly used concept in conventional text processing. Simply put, it means "a group of 'n' consecutive words in a phrase."
  For instance, in the sentence “I like to swim”, we could have n-grams such as:
  - ▪ 1-gram (unigram): “I”, “like”, “to”, “swim”
  - ▪ 2-gram (bigram): “I like”, “like to”, “to swim”
  - ▪ 3-gram (trigram): “I like to”, “like to swim”
  - ▪ 4-gram: “I like to swim”

  Note that the words in an n-gram are taken in order, so “swim I like to” is not a valid 4-gram.

II. AUTHOR’S CONTRIBUTION

In this paper, we aim to investigate the diverse strategies utilized to properly categorize Hinglish and English texts, to compare and contrast the methods used to translate Hinglish text to English, and to identify the research gaps in these methodologies. We then aim to construct an appropriate framework to categorize content as either Hinglish or English and to convert text from Hinglish to English.

III. RELATED WORK

Authors S. H. Attri, T. V. Prasad, and G. Ramakrishna in [5] first determined if the sentence contained an expression or idiom, then extracted it. After tokenization, the phrase was classified as Hinglish, English, or Hindi, depending on its original language. They then used morphological and reverse morphological analysis on each term. POS Tagging sorted the words after analysis and Translation was carried out. "MujhE file send kara as soon as possible" and "asap" were translated into Hindi. The resulting phrase in Hindi was "mujhE yathA shIghra sanchikA bhEj" which translates to "Send me the file as soon as possible" in English. Around 12,000 Hinglish terms were labelled with idioms and translations. Pure Hindi sentences were much more accurate than pure English sentences. Thus, Hinglish is Hindi with English words added, using Hindi syntactic and semantic components instead of English ones. Because of this, it was found that Hinglish sentences translated into pure Hindi were more accurate than those translated into English.

In article [3], D. Gautam, K. Gupta, and M. Shrivastava used fine tune mBART. This approach is a multilingual sequence-to-sequence denoising auto-encoder. Since Precog data includes transliterated Hindi words, they pre-process them in Devanagari using CSNLI (https://github.com/irshadbhat/csnli). They then created a refined variant, the mBART-hien-cm. They transliterated Hinglish terms into Hindi and interpreted the document. Their BLEU score was 33.30.

In [1], the authors deleted sentences with more than 40 tokens, less than 5, or more than 90% or less than 50% OOV (Out-of-vocabulary) terms (so that it can be Hinglish not English with one Hindi word). They then examined the results of Google Translate and Bing Translate, which were 0.139 and 0.14 points, respectively. The scoring system that was suggested by them resulted in a score of 0.153. They labelled each token as Hindi, English, or another language and used Google Translate to convert just the Hindi tokens to English before combining them with their English phrases. PHINC: A Parallel Hinglish social media Code-Mixed Corpus for Machine Translation was used. Annotated corpus data exceeds 50,000 rows.

Authors in [2] propose a four-phase pipeline for automatic Hinglish-to-English translation. They also discuss contextual concerns such as "chalega" which means both "moving" and "will It work?" and compares "code switching" (intersentential) and "code mixing" (intra sentential). This study used no comparable corpus. The language was tagged, transliterated into Devanagari, translated from English to Hindi, combined with Hindi, and translated back into English. A dataset with 25,000 rows was employed, along with LSTM 2-layer, dense layer relu, sigmoid, and adam optimizer. The Loss function used to optimize the method had 0.9 validation accuracy and 0.35 validation loss. The text analysis method used vectorized value -> predict -> tag, along with machine (parallel training data), and value-based (character mapping) transliteration. Google API and RNN were used for Hindi-to-Devanagari back translation. Metrics include the Bi-Lingual Evaluation Understudy (BLEU), Translation Error Rate (TER), and Word Error Rate (WER) score. This method had several drawbacks, including "uss aadmi" becoming "US person" instead of "that person". To solve the spelling variations in the comments and to improve the resource for translation, the authors of article [6] adopt a technique that generates n-best-lists in the training dataset's original language. According to them, a little in-domain dataset could help the translation system perform even better. Our Their methods demonstrate an improvement in translation quality over the baseline system in terms of automatic assessment ratings for the Hindi-English mixed comments that they received from Facebook.

The performance of code-mixed Hinglish to English machine translation jobs is improved by using the large multilingual transformers (mBART and mT5) and fine-tuning them in a dual curriculum learning manner, according to researchers in [7]. They demonstrate a margin of 92.8% above the BLEU scores a very large improvement over the PHINC baseline.

There has recently been research done on code-mixed data, notable ones include language labelling. Using a two-stage technique, Bhattu and Ravi (2016) [8] have classified languages. Character n-grams at the sentence level made up the first level of categorization, while character n-grams at the word level made up the second level. To address the issue of code-mixed translation, Dhar et al. (2018) [9] produced a parallel corpus of code-mixed English-Hindi and English. They have also provided a pipeline for augmentation, which is utilized to improve code-mixed translation efficiency of current machine translation systems.

Through a literature review, many different frameworks that are often used for translating from Hinglish to English were explored. In addition, the challenges that were encountered by these frameworks and approaches have been employed to find solutions to these problems were also studied in great depth.

IV. METHODOLOGY

A. *Method 1 [1]*

1. Preprocessing: The dataset was cleansed of all special characters and numbers, all links were removed using regex, and sentences were trans-formed to lowercase.
2. Google Translate: In this procedure, the source language was set to automatically detect and the destination language was set to English. Using BLEU scores, the results were then analyzed.
3. Proposed Pipeline + Google Translate: In the pipeline proposed by Srivastava and Singh (2020) the text is tokenized first, and then each token is identified as Hindi or English using google detect. The labeled tokens that were labeled as Hindi were translated to English using the Google Trans API and then blended with the rest of the tokens to construct whole sentences.

Challenge: Through this framework it was found that the context of the Hinglish word in the English sentence was not established clearly on translation. Another challenge that was discovered in this method was that, google detect was not able to effectively classify tokens as Hindi or English. For instance, the word “uss” in Hindi would mean “that”, however google detect would interpret it as “us”.

Solution: One solution to this problem would be to design a classification model that has been trained on Hinglish and English words.

B. *Method 2 [2]*

1. Classification Model: Dataset: The dataset used to construct the classification model was made public by Precog [5] and comprised 25,000 of the most popular Hindi and English terms. Preprocessing: Using a Label Encoder, each word in the dataset was tagged as either “hi” or “en”, and these labels were encoded as 0 and 1. The words in the dataset were first converted to lowercase and then vectorized. Model Training: The model used was a 2-layer LSTM with relu and sigmoid activation functions and an Adam optimizer. The validation accuracy and validation loss of the model were 91% and 0.422 respectively.
2. Machine Transliteration & Translation of Tokens: The classification model was used to categorize each token as either Hindi or English. The tokens labeled "hi" were transliterated from Hinglish to Devanagari Hin-di using the Indic Transliteration package, whilst the tokens labeled "en" were translated into their Hindi equivalents using the Google Trans API. Afterwards, these tokens were combined to form complete phrases.
3. Machine Translation from Hindi to English: Google translate was used to convert the Devanagari Hindi sentences into English. The translations were compared and assessed using the BLEU score. The translation of this framework was found to be better than the previous framework, Nonetheless, several spelling and grammatical problems were discovered during the transliteration of Devanagari Hindi to English; the following framework seeks to reduce these inaccuracies.

***Article Received**: 26 April 2023 **Revised**: 20 June 2023 **Accepted**: 01 July 2023*

V. PROPOSED METHODOLOGY

It was found that a lot of sentences after getting translated had spelling and grammatical errors. The proposed architecture as shown in Fig 1. adds an extra layer for spelling and grammar correction to improve accuracy.

The Devanagari Hindi translations and transliterations featured numerous misspellings and grammatical errors. Hence, the translations received a low BLEU score. In order to address this issue, the Spello model was employed to correct the spellings of the translated Devanagari Hindi words. This model uses phonetic principles to identify words with comparable sounds.

Example:

Incorrect Spelling 1: अच्वा
Correct Spelling 2: अच्छा

Gramformer is a model that identifies grammatical errors in sentences and re-places them with their grammatically correct alternatives in order to address the issue of bad sentence grammar.

Example:

Incorrect Sentence 1: We likes Pizza
Correct Sentence 1: We like Pizza
Incorrect Sentence 2: How is you?
Correct Sentence 2: How are you?

VI. RESULTS

A. Dataset

We used two code-mixing datasets released by Dhar et al. (2018) [9] and Srivastava and Singh (2020) [1] for this work. These datasets were chosen for their variety of sources (all major social networking platforms), which reduces bias in the dataset. The dataset produced by Srivastava and Singh (2020) [1] contains 13,760 rows of Hinglish social media comments and tweets transformed to English counterparts. This dataset consists of 103,887 Hinglish tokens and 96,439 English and other tokens. Table 1 has the first five entries of the dataset that comprise the Hinglish sentence and its English translation.

Dhar et al. (2018) [9] developed a dataset for machine translation of code-mixed data that included 6,096 English-Hindi code-mixed and English monolingual gold standard parallel sentences.

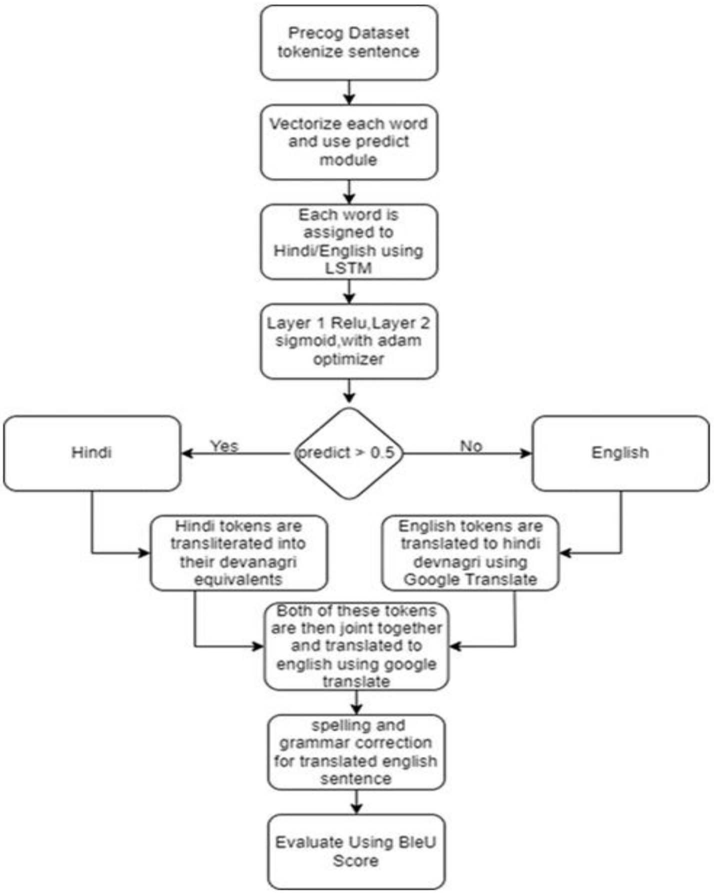
Precog Dataset tokenize sentence
Vectorize each word and use predict module
Each word is assigned to Hindi/English using LSTM
Layer 1 Relu,Layer 2 sigmoid,with adam optimizer
predict > 0.5
Yes
No
Hindi
English
Hindi tokens are transliterated into their devanagri equivalents
English tokens are translated to hindi devnagri using Google Translate
Both of these tokens are then joint together and translated to english using google translate
spelling and grammar correction for translated english sentence
Evaluate Using BleU Score

Figure 1. Proposed Architecture

TABLE I. TRANSLATION USING PROPOSED MODEL ON DATASET IN [1]

Index	Hinglish	English
0	@someUSER congratulations on your celebrating british kid singers sophia grace's and rosie's 1st anniversary of a visit of your show.	@some users congratulate you for celebrating British kid singers Sophia Grace's and Rosie's 1st anniversary visit of your show
1	@LoKarDi_RT uske liye toh bahot kuch karna padega ye pappiyon se kaam nahi chalega #ForTheSakeOfHumanity	@Lokardi_ rat we should a lot more for that, by this evi people nothing will happen #ForTheSakeOfHumanity
2	@slimswamy yehi to hum semjhane ki koshish kar rahe hain. Log to sab kuch ko issi mein tol dete hain.	@Slimswam, this is what I'm expecting you to understand, people invest everything in this isn't it.
3	@DramebaazKudi cake kaha hai??	@Where is Dramebajakudi where is the cake?
4	@Vyomnaut sahi me yaar	@Vyomnaut right man

This dataset was used for machine translation of code-mixed data. There are a total of 37,673 Hinglish tokens and 26,276 other tokens, including English to-kens contained within this dataset. The Hinglish statement and its translation in English are presented in the following format across the first five rows of the dataset in Table 2.

TABLE II. TRANSLATION USING PROPOSED MODEL ON DATASET IN [9]

Index	Hinglish	English
0	Apne aqirat ki fiqar karo wanha kon bachega sirf allha tobakarle nada	worry about your aqirat who will be left there only God
1	apka fan ho bangladesh me plz cal sallu bhai met u +8801719447771 cal bhai	I am your fan in Bangladesh please call sallu bhai met you +8801719447771 call bhai
2	Flop jaaigi movie teri.... aehsan framosh..	your movie will be a flop .... ungrateful...
3	Kya socha hai shaadi k bare mai	what have you thought about marriage
4	Mami papa, aur bacha party aur Sab kaise hai	mother father and kids party how is everyone else

The dataset that was utilized in the construction of the classification model was made available to the public by Precog [5]. It contained 25,000 of the most common words in both Hindi and English. The displayed portion of data includes two columns: one for the word, and the other for the label that corresponds to it as shown in Table 3.

TABLE III. WORDS AND THEIR LABEL

Index	Word	Label
0	Cricket	En
1	Chal	Hi
2	Raha	Hi
3	Hain	Hi
4	Yaha	Hi

Summary of results for all the approaches have been mentioned in Table 4.

TABLE IV. RESULT SUMMARY

Method Name	Method Description	Dataset 1 (Srivastava and Singh 2020)	Dataset 2 (Dhar 2018)	Learnings
Method 1 [1]	Google Translate	58.1	64.8	Google API does not allow more than 10,000 calls
Method 1.5 [1]	Used Google Translate to only translate Hindi tokens and embed them into the English sentence	46.4	46.2	The context of the Hinglish token in the English sentence is not established
Method 2 [2]	LSTM model to detect a token as "hi" or "en". Tokens labeled "en" are translated to Hindi and tokens labeled "hi" are transliterated to Devanagari Hindi. The whole Hindi sentence is then translated to English.	46.8	48.6	The google detect function couldn't identify Hindi words from English words. Transliteration helps correct the Hinglish spellings and gives it uniformity.
Method 3 [Proposed Model]	Same method as above, but contains a spelling and grammar correction module.	48.2	51.4	Corrects the grammar of the translated

VII. CHALLENGES

- Hindi is Context Dependent:
  Many Hinglish terms have several meanings that can be determined from the sentence context alone [1]. An example of this could be the term “chalega” which in some sentences means “will work” and in some sentences means “walk”
- Lack of Standardization:
  There are no standard spellings in Hinglish; most users rely on the phonetics of the word to determine its Romanized spelling [1], thus resulting in a variety of words with the same meaning but different spellings. For instance, “Nahi”, “Nai”, “Nhi” all mean “No” in English. Another ex-ample could be “main”, “mai”, “mein” all meaning “me” in English.
- Unstructured Data:
  Hinglish is most frequently encountered on informal platforms. Since informal writing rarely adheres to punctuation, correct spelling, and correct grammar, this adds an additional layer of complication to the translations. In addition, the use of abbreviated words might lead to misunderstanding on whether a word is Hindi or English.
- Similar words in Hindi and English:
  Several words used in Hinglish are also found in English [2], in such cases it creates an obstacle for the model to determine the source language and thus choose between transliterations and translations. For in-stance, the term “main” can be used in both Hinglish where it means “me” and English where it means “principal”.
  Figure 2 shows an instance of confusion between an English and Hindi token.

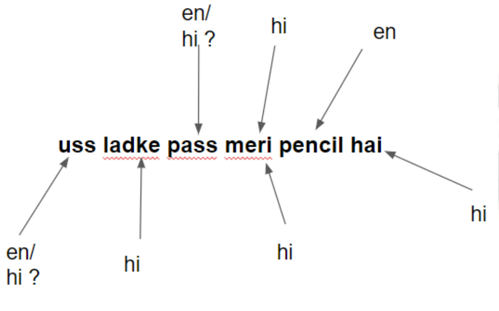

Figure 2. Token Labelling

- Lack of Processing Tools, Packages and Techniques:
  There is a severe lack of research and large-scale technologies that can be used to efficiently produce POS tags, Named Entity Recognition, or even Word bedding because much of the research on NLP is done mostly in English [10], with some of it in monolingual languages. This makes it difficult to design large-scale algorithms that can analyze, classify, or determine the sentiment of Hinglish text, among other things.

- Availability of Data:
  Since Hinglish is typically spoken in casual conversational contexts, the vast majority of relevant data is not publicly available and hence cannot be acquired [10]; yet, its availability would greatly facilitate the resolution of the issues discussed.

VIII. CONCLUSION & FUTURE WORK

This paper presents an implementation of code-mixed translation strategies for Hinglish to English translation. Our learnings from the implementations have led us to develop a framework that represents an improvement over previous method, although it may not yet surpass Google's Translation API. Furthermore, this research highlights several obstacles that must be overcome to develop truly effective code-mixed machine translation systems. The insights gained from this study can be extended to other code-mixed text, expanding the potential applications of our work.

In the future, our research may involve training machine translation models such as mBART and mT5 on the datasets we have created. However, one major challenge remains the scarcity of code-mixed Hinglish data [3]. To address this, future research could focus on generating larger parallel corpuses [6]. Additionally, there is a need for pre-trained code-mixed translation models, as training models like mBART from scratch is computationally demanding.

Therefore, this study improves the understanding of code-mixed machine translation and proposes research topics that will facilitate the creation of more robust translation systems.

REFERENCES

[1] Srivastava, V., Singh, M.: PHINC: A Parallel Hinglish Social Media Code-Mixed Corpus for Machine Translation. (2020).

[2] Jadhav, I., Kanade, A., Waghmare, V., Chandok, S.S., Jarali, A.: Code-Mixed Hinglish to English Language Translation Framework. In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). pp. 684–688.IEEE(2022). https://doi.org/10.1109/ICSCDS53736.2022.9760834.

[3] Gautam, D., Gupta, K., Shrivastava, M.: Translate and Classify: Improving Sequence Level Classification for English-Hindi Code-Mixed Data. In: Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching. pp. 15–25. Association for Computational Linguistics, Stroudsburg, PA, USA (2021). https://doi.org/10.18653/v1/2021.calcs-1.3.

[4] Dhawal Khem, Shailesh Panchal, Chetan Bhatt. (2023). Text Simplification Improves Text Translation from Gujarati Regional Language to English: An Experimental Study. International

Journal of Intelligent Systems and Applications in Engineering, 11(2s), 316–327. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2699

[5] Chakrawarti, R.K., Bansal, P.: Approaches for Improving Hindi to English Machine Translation System. Indian J Sci Technol. 10, 1–8 (2017). https://doi.org/10.17485/ijst/2017/v10i16/111895.

[6] Attri, S.H., Prasad, T.V., Ramakrishna, G.: HiPHET: A Hybrid Approach to Translate Code Mixed Language (Hinglish) to Pure Languages (Hindi and English). Computer Science. 21, (2020). https://doi.org/10.7494/csci.2020.21.3.3624.

[7] Dr. B. Maruthi Shankar. (2019). Neural Network Based Hurdle Avoidance System for Smart Vehicles. International Journal of New Practices in Management and Engineering, 8(04), 01 - 07. https://doi.org/10.17762/ijnpme.v8i04.79

[8] Singh, T.D., Solorio, T.: Towards Translating Mixed-Code Comments from Social Media. In: International Conference on Computational Linguistics and Intelligent Text Processing. pp. 457–468 (2018). https://doi.org/10.1007/978-3-319-77116-8_34.

[9] Agarwal Vibhav, Rao Pooja, Jayagopi Dinesh Babu: Hinglish to English Machine Translation using Multilingual Transformers. Proceedings of the Student Research Workshop Associated with RANLP 2021. 16–21 (2021).

[10] Martin, S., Wood, T., Hernandez, M., González, F., & Rodríguez, D. Machine Learning for Personalized Advertising and Recommendation. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/156

[11] Sristy, N.B., Krishna, N.S., Krishna, B.S., Ravi, V.: Language Identification in Mixed Script. In: Proceedings of the 9th Annual Meeting of the Forum for Information Retrieval Evaluation. pp. 14–20. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3158354.3158357.

[12] Dhar Mrinal, Kumar Vaibhav, Shrivastava Manish: Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach. Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing. 131–140 (2018).

[13] Srivastava, V., Singh, M.: Challenges and Considerations with Code-Mixed NLP for Multilingual Societies. (2021).