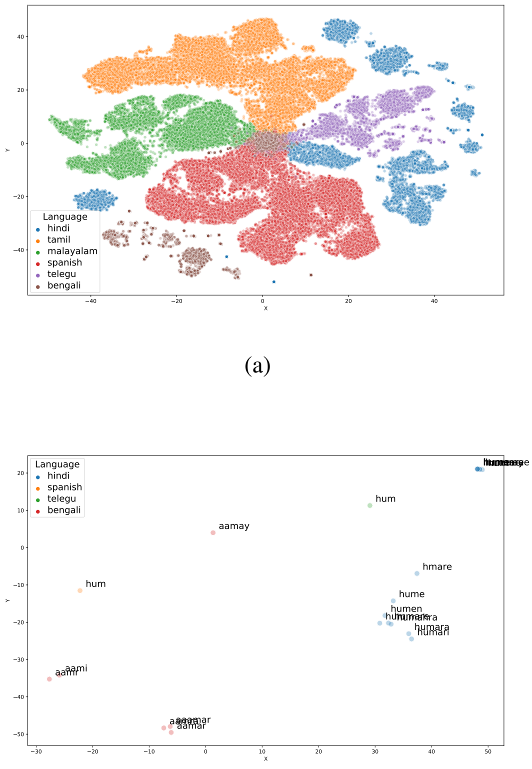
# HIT: A Hierarchically Fused Deep Attention Network for Robust Code-mixed Language Representation (Appendix)

**Ayan Sengupta[1], Sourabh Kumar Bhattacharjee[1],**
**Tanmoy Chakraborty[2], Md Shad Akhtar[2]**
[1]Optum Global Advantage (UnitedHealth Group), Noida, India
[2]Dept. of CSE, IIIT-Delhi, India
{ayan.sengupta007, sourabhb398}@gmail.com;
{tanmoy, shad.akhtar}@iiitd.ac.in

## 1 Semantic Understanding of Languages



(a)

(b)

Figure 1: t-SNE visualization (a) of all words; (b) of selected pronouns. Overlapping clusters show how semantically similar words from different languages are embedded onto a similar space.

In this section, we study the semantic relationships between different Indic languages. We calculate the proportion of common words in Table 1 between different language pairs to understand the multilingualism in India. We observe that Bengali code-mixed texts have the highest proportion of English words 32% as compared to other languages. Moreover, 50% of all Bengali words are also present in the Hindi CM texts, although 58% of those words are English. We observe that users using Hindi CM texts use very few words taken from other languages. On the other hand, a significant proportion of Bengali and Telugu CM words are common in other languages, although, majority of them are English. The two Dravidian languages, Tamil and Malayalam, show a very distinctive behavior. They share very little linguistic similarity with other Indic languages. On the other hand, 10% of all Tamil words are used in Malayalam and 17% of all Malayalam words are used in Tamil. Moreover, this sharing is not driven by English, as, only 27% of these words are English, which is the lowest proportion among all other language pairs. Being originate from a similar root and having a phonetic resemblance makes Tamil and Malayalam *sister languages*[1]. Similar observations are also made from the word representation lens. We use t-SNE (Van der Maaten and Hinton, 2008) plots to embed HIT's representations onto a 2-D space for interpretability (Fig 1). Although, the embeddings are well clustered based on the languages, we can easily figure out the semantically similar words across languages embedded onto a similar space. Furthermore, Fig 1(b) shows that pronouns (e.g., '*aap*') in Tamil, Telegu and Hindi are embedded onto a similar space with Bengali words '*aamar*', '*aamay*'. Although each of these representations are learned on separate models on separate datasets, the robustness of the underlying hierarchical representation enables our model to capture cross-lingual semantics from noisy code-mixed texts. We can attribute these observations to the relatedness of Indic languages on a socio-cultural basis.

[1] https://royalsocietypublishing.org/doi/10.1098/rsos.171504

|  |  | Target Language |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | **Hindi (English)** | **Malayalam (English)** | **Tamil (English)** | **Bengali (English)** | **Telegu (English)** | **Spanish (English)** |
| **Source Language** | Hindi | 1.00 (0.16) | 0.02 (0.41) | 0.04 (0.39) | 0.02 (0.58) | 0.02 (0.57) | 0.07 (0.62) |
|  | Malayalam | 0.14 (0.41) | 1.00 (0.06) | 0.17 (0.27) | 0.03 (0.71) | 0.05 (0.57) | 0.07 (0.64) |
|  | Tamil | 0.15 (0.39) | 0.10 (0.27) | 1.00 (0.07) | 0.03 (0.69) | 0.05 (0.56) | 0.07 (0.64) |
|  | Bengali | 0.50 (0.58) | 0.16 (0.58) | 0.23 (0.69) | 1.00 (0.32) | 0.21 (0.71) | 0.36 (0.72) |
|  | Telegu | 0.36 (0.57) | 0.15 (0.57) | 0.29 (0.56) | 0.12 (0.71) | 1.00 (0.22) | 0.28 (0.65) |
|  | Spanish | 0.12 (0.62) | 0.02 (0.64) | 0.03 (0.64) | 0.02 (0.72) | 0.03 (0.65) | 1.00 (0.11) |

Table 1: Proportion of words in source language in the target language.

| Lang | POS tags |
|---|---|
| Hindi (14) | *X, VERB, NOUN, ADP, PROPN, ADJ, PART, PRON, DET, ADV, CONJ, PART_NEG, PRON_WH, NUM* |
| Bengali (39) | *N_NN, V_VM, RD_PUNC, N_NNP, PSP, PR_PRP, JJ, RB_AMN, CC, QT_QTF, DM_DMD, RP_RPD, @, RD_RDF, V_VAUX, DT, PR_PRQ, #, RP_NEG, E, $, RB_ALC, N_NNV, PR_PRL, N_NST, RP_INJ, RD_SYM, DM_DMR, RP_INTF, PR_PRF, DM_DMQ, QT_QTO, U, QT_QTC, PR_PRC, RD_ECH, QY_QTO, Ã°Å¸Ëœ, ∼* |
| Telugu (52) | *N_NN, N_NNP, RD_RDF, RD_PUNC, V_VM, JJ, @, PSP, PR_PRP, RP_INJ, DT, RB_AMN, CC, $, U, E, #, N_NNV, &, PR_PRQ, V_VAUX, RD_PUNC", ∼, RD_RDFP, QT_QTF, RD_UNK, DM_DMD, RP_RPD, RB_ALC, DM_DMQ, RD_ECH, N_NST, acro, PR_PRL, QT_QFC, RP_RDF, PR_PRC, r, RD_SYM, RD_RDFF, psp, PR_PRF, QT_QTP, RD_P/UNC, PR_PPR, PR_RPQ, RPR_PRP, RP_INTF, -* |
| Spanish (17) | *VERB, PUNCT, PRON, NOUN, DET, ADV, ADP, INTJ, CONJ, ADJ, AUX, SCONJ, PART, PROPN, NUM, UNK, X* |

Table 2: POS tagsets for different datasets.

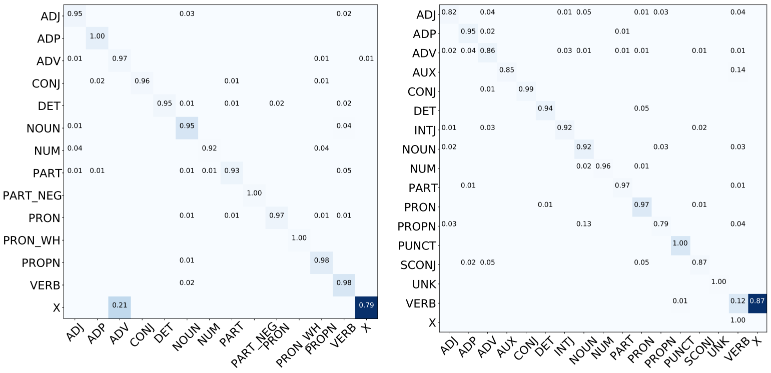
## 2 Datasets

We report all available POS tags in Table 2.

## 3 Confusion Matrices and Error Analysis

We report the confusion matrices to show the label-wise performance for the sentiment classification, PoS tagging and NER in Tables 3, 2, and 4, respectively.

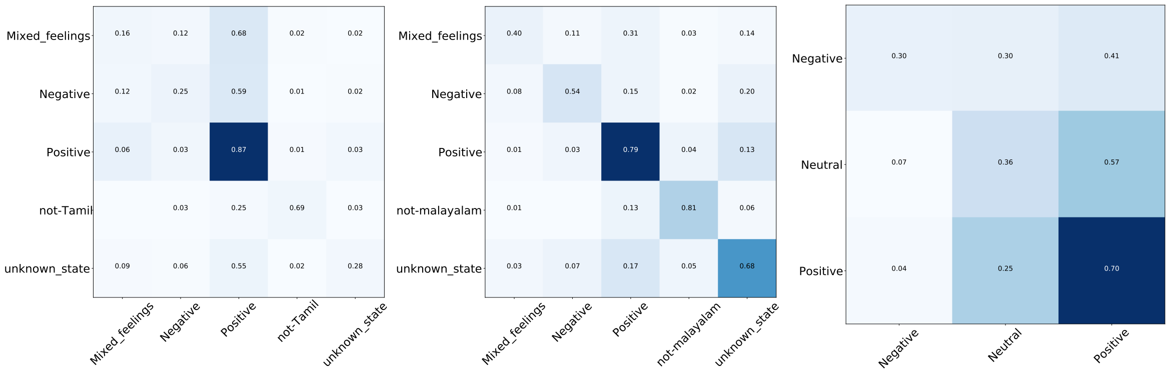
We similarly perform qualitative analysis on the MT task where our model shows superior performance as compared to the baselines. In example 1 of Table 3 (d), HIT translates the English text ‘‘*Licencing and import policies were liberalise”* to *“Licencing aur policies liberal the* |*”*. Although this prediction has very low BLEU score when evaluated against the target, this example shows an interesting observation. The overall translation is a contextually meaningful sentence in Hindi. Further HIT translates the phrase *‘were liberalise’* to *‘liberal the’*. In Hindi, *‘the’* represents past tense. Another interesting observation is the ability of HIT to copy texts from source to predicted text. Even without having an explicit copying mechanism (See et al., 2017), HIT is able to understand the key phrases that co-occur in both Hindi and English, like, numeric and proper nouns and copies these tokens while generating. This shows how our model can also be used in conditional generation of texts. It also ends the sentence with |, which is a common punctuation widely used as a full stop in Hindi texts.



(a) Hindi

(b) Spanish

Figure 2: Confusion matrices of HIT on POS tasks. Due to high cardinality of output classes, we do not report for Bengali and Telugu.

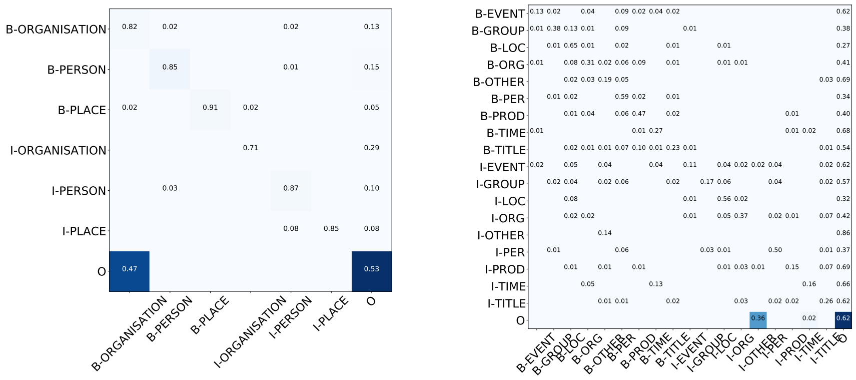


(a) Tamil

(b) Malayalam

(c) Spanish

Figure 3: Confusion matrices of HIT on sentiment tasks.



(a) Hindi

(b) Spanish

Figure 4: Confusion matrices of HIT on NER.

## References

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

|  | **Input** | Sys | **Prediction** |
|---|---|---|---|
| 1 | **Org:** *#surgicalstrike_X #pakistan_X will_V not_Neg sleep_N in_ADP peace_N tonight_N ._X khamoshi_V toofan_V ke_ADP aane_V ki_ADP aahat_N to_P nahi_Neg* | A | *#surgicalstrike_X #pakistan_X will_V not_part_Neg sleep_N in_ADP peace_N tonight_N ._X, khamoshi_V toofan_V ke_ADP aane_V ki_ADP aahat_N to_P nahi_part_neg* |
|  | **Translated:** *#surgicalstrike #pakistan will not sleep in peace tonight. Does this silence signify that a storm is approaching* | B | *#surgicalstrike_X #pakistan_X will_V not_part_Neg sleep_N in_ADP peace_N tonight_N ._X, khamoshi_V toofan_V ke_N aane_V ki_ADP aahat_ADP to_P nahi_part_Neg* |
| 2 | **Org:** *minimum_N cincuenta_Num mil_Num por_ADP persona_N ._Punct* | A | *minimum_N cincuenta_Num mil_Num por_ADP persona_N ._Punct* |
|  | **Translated:** *minimum fifty thousand per person .* | B | *minimum_N cincuenta_Num mil_Num por_ADP persona_N ._Punct* |

(a) PoS

|  | **Input** | Sys | **Prediction** |
|---|---|---|---|
| 1 | **Org:** *@gurmeetramrahim {dhan dhan satguru}$_{Per}$ tera hi aasra #msgloveshumanity salute 2 {msg}$_{Org}$ <url>* | A | *@gurmeetramrahim {dhan dhan satguru}$_{Per}$ tera hi aasra #msgloveshumanity salute 2 msg <url>* |
|  | **Translated:** | B | *@gurmeetramrahim {dhan dhan satguru}$_{Per}$ tera hi aasra #msgloveshumanity salute 2 {msg}$_{Org}$ <url>* |
| 2 | **Org:** *ste {sábado}$_{Time}$ nuestras alumnas en {imagen modeling}$_{Org}$ by {la gatita}$_{Per}$ reciben la visita de {monic abbad}$_{Per}$ , joven … <url>* | A | *ste {sábado}$_{Time}$ nuestras alumnas en imagen modeling by la gatita reciben la visita de {monic}$_{Per}$ abbad , joven … <url>* |
|  | **Translated:** *This saturday our students in image modeling by the kitten receive a visit from young monic abbad* | B | *ste {sábado}$_{Time}$ nuestras alumnas en imagen modeling by la gatita reciben la visita de {monic abbad}$_{Per}$ , joven … <url>* |

(b) NER

|  | **Input** | **Gold** | **Prediction** |  |
|---|---|---|---|---|
|  |  |  | **A** | **B** |
| 1 | **Org:** *safal videsh yatra ke liye badhai ho sir*<br>**Trans:** *Congratulations on the successful foreign trip sir* | Pos | Pos | Neu |
| 2 | **Org:** *nunca pensé que " bruh " me frustraría tanto*<br>**Trans:** *I never thought that "bruh" would frustrate me so much* | Neu | Neu | Neg |
| 3 | **Org:** *desh chodo pahaley yeh media ko change karo ... !! ?*<br>**Trans:** *Leave the country, first change the media* | Neu | Neg | Neg |

(c) Sentiment

|  |  |
|---|---|
| 1 | **Source:** *Licencing and import policies were liberalise*<br>**Reference:** *license tatha import ki policies ko udar banaya gaya* |
|  | **HIT:** Licencing aur policies liberal the । |
| 2 | **Source:** *This fact is based on possibility*<br>**Reference** *yah fact possibility par aadharit hai ।* |
|  | **HIT:** yah fact possibility par aadharit hai |

(d) MT

Table 3: Error Analysis. System A denotes HIT and B denotes CS-ELMO.