IndiSocialFT: Multilingual Word Representation for Indian languages in code-mixed environment

Saurabh Kumar, Sanasam Ranbir Singh and Sukumar Nandi
Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati, Assam, India
{saurabh1003, ranbir, sukumar}@iitg.ac.in

1. Introduction

The increasing number of Indian language users on the internet necessitates the development of Indian language technologies. In response to this demand, this paper presents a generalized representation vector for diverse text characteristics, including native scripts, transliterated text, multilingual, code-mixed, and social media-related attributes. Text from both social media and well-formed sources is gathered and the FastText model is utilized to create the "IndiSocialFT" embedding. Through intrinsic and extrinsic evaluation methods, IndiSocialFT is compared with three popular pre-trained embeddings trained over Indian languages. Findings show that the proposed embedding surpasses the baselines in most cases and languages, demonstrating its suitability for various NLP applications.

2. Available pre-trained embedding models for Indian Languages

	Native	Multilingual	Code-mixed	#lang	#tokens	Sources
FT-WC	✓	–	–	17	–	Common Crawl and Wiki
IndicFT	✓	–	–	11	8.8 B	News Crawls
IndicBERT	✓	✓	–	12	8.8 B	News Crawls
MuRIL	✓	✓	✓	16	11.0 B	OSCAR corpus, Wiki, PMIndia, Dakshina
IndiSocialFT	✓	✓	✓	20+	11.0 B	Social Media, Samanantar and Dakshina Dataset, Wiki

Summarization of different model support and corresponding training dataset

3. Proposed Dataset and Model

Data Collection:

- Crawled Twitter using Twitter's API, focusing on Indian Location from 2019 to 2022.
  – This resulted in a total of 0.6 billion tweets, equivalent to 5.5 billion tokens.
- Collected posts and comments from Facebook profiles of well-known Indian individuals.
  – A total of 0.8 million posts, including comments and nested comments, resulting in 14.8 million tokens.
- Extracted comments on videos uploaded by popular news and entertainment channels on YouTube.
  – Gathered 0.4 million comments from YouTube, comprising 3.8 million tokens.
- To ensure balanced distribution included 20 Indian languages in native scripts from various public datasets
  – Added 0.3 billion sentences to the social media dataset, comprising 5.3 billion tokens.
- **Language Support :** Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pn), Tamil (ta), Telugu (te), Sindhi (sd), Sinhala (si), Urdu (ur), Manipuri (mni), Sanskrit (sa), Bhojpuri (bho), Nepali (ne), Maithili (mai), and Angika (ang)

Proposed Embedding Model :

- Trained a 300-dimensional embeddings model using FastText
- Run the training for 15 epochs, utilized a window size of 5, and set a minimum token count of 5 for each instance

4. Evaluation Methodology

Intrinsic Evaluation (with Native Script):

- Ranking-based intrinsic evaluation
- Word similarity based intrinsic evaluation using the IIIT-Hyderabad word similarity dataset

Extrinsic Evaluation (with both Native and Code-mixed):

- Adopted the k-NN (k = 4) classifier based evaluation method.
- Utilized IndicGLUE Datasets, Dravidian-CodeMix-FIRE 2021 dataset, YouTube cookery channels viewer comments in Hinglish, and Hinglish-TOP Dataset for evaluating performance on text classification task.

5. Evaluation on Texts in Native Scripts

Intrinsic Evaluation:

Lang	FT-WC	IndicFT	IndiSocialFT
pa	0.384	0.445	**0.683**
hi	0.551	0.598	**0.664**
gu	0.521	0.600	**0.665**
mr	0.544	0.509	**0.624**
te	0.543	0.578	**0.662**
ta	0.438	0.422	**0.691**
ur	0.248	NA	**0.624**
Avg	0.461	0.525	**0.659**

Word Similarity results for different pre-trained embeddings. (a) FT-WC, (b) IndicFT, (c) IndiSocialFT .

Extrinsic Evaluation:

Lang	FT-WC	IndicFT	IndiSocialFT
pa	95.53	**96.47**	95.51
bn	97.57	**97.71**	97.14
or	96.20	**98.43**	97.23
gu	94.63	99.02	**99.51**
mr	97.07	**99.37**	98.74
kn	96.53	**97.43**	96.36
te	98.08	**99.17**	99.04
ml	89.18	**92.83**	90.50
ta	95.90	**97.26**	96.20
Avg	95.63	**97.52**	96.69

Accuracy score (in percentage) on IndicGLUE dataset

6. Evaluation on Multilingual Code-Mixed Texts

Dataset	TF-IDF		MuRIL		IndicBERT		IndiSocialFT	
	acc	F1	acc	F1	acc	F1	acc	F1
hi-en(YT)	0.579	0.551	0.652	0.641	0.606	0.591	**0.661**	**0.661**
hi-en(TOP)	0.839	0.836	0.864	0.867	0.764	0.758	**0.922**	**0.912**
ml-en(SA)	0.144	0.068	0.531	**0.465**	0.510	0.410	**0.539**	0.463
ml-en(OfD)	0.893	0.315	0.925	**0.398**	0.923	0.384	**0.926**	0.389
ta-en(SA)	**0.579**	0.208	0.564	0.421	0.538	0.374	0.528	**0.427**
ta-en(OfD)	0.731	0.176	0.734	0.349	0.725	0.321	**0.740**	**0.381**
kn-en(SA)	0.487	0.288	0.546	**0.440**	0.522	0.409	**0.556**	0.427
kn-en(OfD)	0.617	0.222	**0.677**	0.361	0.656	0.320	0.652	**0.368**
Average	0.609	0.333	0.686	0.493	0.656	0.446	**0.691**	**0.504**

Accuracy (acc) and Macro-F1 (F1) score of Text Classification task on different code-mixed dataset

7. Conclusion and Future Work

- Addressed the challenge of representing text in a multilingual code-mixed social media environment by developing a FastText-based embedding model
- In future work, it is planned to further improve the quality of the proposed embeddings by incorporating additional data sources and exploring transformer-based pre-training techniques.