IndiSentiment140: Sentiment Analysis Dataset for Indian Languages with Emphasis on Low-Resource Languages using Machine Translation

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

Abstract

Sentiment analysis, a fundamental aspect of Natural Language Processing (NLP), involves the classification of emotions, opinions, and attitudes in text data. In the context of India, with its vast linguistic diversity and low-resource languages, the challenge is to support sentiment analysis in numerous Indian languages. This study explores the use of machine translation to bridge this gap. The investigation examines the feasibility of machine translation for creating sentiment analysis datasets in 22 Indian languages. Google Translate, with its extensive language support, is employed for this purpose in translating the Sentiment140 dataset. The study aims to provide insights into the practicality of using machine translation in the context of India's linguistic diversity for sentiment analysis datasets. Our findings indicate that a dataset generated using Google Translate has the potential to serve as a foundational framework for tackling the low-resource challenges commonly encountered in sentiment analysis for Indian languages.

1 Introduction

Within the domain of natural language processing (NLP) research, a notable challenge is the scarce or limited availability of resources, particularly in the context of Indian languages. This scarcity extends to the field of sentiment analysis as well. As NLP researchers and practitioners continue to explore innovative solutions to bridge the gap between resource-rich languages like English and resource-poor languages in the Indian subcontinent, this study embarks on a critical inquiry into one such challenge: the support for Low-resource languages within the Indian context, specifically in the realm of sentiment analysis.

A fundamental realization emerges – fulfilling the resource requirements for Indian languages necessitates the creation of these resources. This entails manually creating linguistic datasets, including labeled text, grammatical structures, and socio-cultural and language specific sentiment indicators. While this approach is undoubtedly effective, it raises questions about its feasibility, especially when dealing with the extensive linguistic and cultural diversity found in India. Given the multitude of languages spoken in India, manual resource creation for each language becomes a monumental and perhaps unfeasible task. Considering the experience of the NLP researcher in the context of resource constrain, *is it possible to supplement the resource gap by bootstrapping with partially correct English to Indian languages machine translation tools such as Google Translate?* Various machine translation tools and APIs, including Google Translate, DeepL, Microsoft Translator, and OpenAI's Machine Translation, are readily available. According to the documentation available for Google Translate (Bapna et al., 2022), it supports over 1000+ languages, among which 22 are Indian languages.

Motivated by this, and with the objective of assessing the feasibility and efficacy of machine translation in the context of India's linguistic diversity, particularly in the development of sentiment analysis datasets for Indian languages, this paper considers Sentiment140 (Go et al., 2009) which is considered to be one of the largest corpora for sentiment analysis task in general domain curated from Twitter (currently X). Our selection of the Sentiment140 dataset is driven by its reputation as a significant dataset for sentiment analysis in noisy environments. Its substantial size, balanced distribution, and the inherent challenges of tweets make it an ideal choice for evaluating sentiment analysis models in real-world scenarios. By incorporating this dataset into our study, we address resource constraints in Indian languages while benefiting from a well-established benchmark for assessing sentiment analysis models in noisy environments.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7689–7698

Leveraging this widely used Sentiment140 dataset, we generate parallel corpora, known as *IndiSentiment140*, by translating it into 22 Indian languages supported by Google Translate. We utilize these generated parallel corpora to investigate their suitability for sentiment analysis by training two foundational neural models: the multi-layer perceptron (MLP) and the Long Short-Term Memory (LSTM) models. Despite the availability of more advanced neural and transformer-based models, our choice of these simpler models is driven by the aim to assess whether the translated dataset preserves both the sentiment characteristics of the original dataset and maintains the underlying sequential structure rather than the model aspect. If the dataset is suitable for these two models, it is likely to be suitable for advanced models.

Considering the socio-cultural and language-specific nuances influencing sentiment, we not only leverage this translated corpus but also utilize a human-translated sentiment dataset, indicSentiment (Doddapaneni et al., 2023), to evaluate the sentiment analysis model trained on the translated corpus, *IndiSentiment140*. Our findings indicate that these translated datasets not only retain the sentiment characteristics of the original dataset but also maintain the sequential structure within the translated text across specific languages. This suggests that these translated datasets have the potential to serve as a foundational framework for addressing the low-resource issue that prevails in Indian languages, offering a promising avenue for further NLP research and development in this context. The *IndiSentiment140* dataset is available on https://www.iitg.ac.in/cseweb/osint/resourcess.php.

Lang	Flores	PMI	Indic	Google
en → as	2.955	2.408	14.306	9.6
en → bho	8.645	-	-	16.7
en → bn	14.481	2.493	29.445	-
en → doi	-	-	-	15.3
en → gom	-	-	-	10.8
en → gu	18.629	10.163	35.926	-
en → hi	32.072	20.003	34.001	-
en → kn	15.454	6.356	29.845	-
en → lus	7.230	-	-	13.2
en → mai	5.397	-	-	8.6
en → ml	12.172	0.781	21.961	-
en → mni	-	-	-	12.9
en → mr	12.065	7.579	7.746	-
en → ne	13.655	-	-	-
en → or	15.516	4.865	30.084	-
en → pa	21.175	16.881	34.531	-
en → sa	0.321	-	-	2.8
en → sd	18.273	-	-	-
en → si	1.916	-	-	-
en → ta	12.543	2.194	23.922	-
en → te	17.344	1.229	17.092	-
en → ur	3.230	17.198	39.102	-

Table 1: Performance of Google Translate in terms of BLEU score on different parallel corpora. Here **Lang** column represents the language pair that has been used as the source and the destination language while translating the text. **Flores**, **PMI**, and **Indic** represent the BLEU score of the Google-translated dataset and human-translated dataset taken from Flores-200 dataset (Costa-jussà et al., 2022), PMIndia (Haddow and Kirefu, 2020), and IndicSentiment (Doddapaneni et al., 2023) respectively. The BLEU score in **Google** column is directly taken from the report (Bapna et al., 2022).

2 Related Work

Sentiment analysis in the context of Indian languages presents a unique challenge due to the limited availability of resources. Addressing this low-resource challenge, several studies have introduced sentiment datasets for Indian languages. Most of the existing sentiment analysis datasets are primarily designed for resource-rich languages like Hindi, Bangla, Tamil, Telugu, Malayalam, among a few others. These datasets are typically small in size and rely on human annotations. Here are some of the datasets specifically crafted for sentiment analysis in the context of Indian languages.

A sentiment dataset consists of 8K review sentence in Hindi was introduced by Akhtar et al. (2016). These reviews were sourced from various newspapers, blogs, and e-commerce websites and manually labeled with sentiment. A similar Hindi dataset named BHAAV was introduced by Kumar et al. (2019) , consisting of 20K sentences collected from different short stories.

A manually annotated Telugu dataset, Sentiraama, for sentiment analysis in Telugu, was presented by Gangula and Mamidi (2018). This dataset integrates multiple domain sources to improve sentiment prediction, comprising 1K documents from various domains, including song lyrics, movie reviews, product reviews, and book reviews.

In a similar vein, Md. Rezaul Karim and Cochez (2020) presented a dataset of 320K documents

Lang	#sample	#min	#max	#avg	%eng	#train	#test
as	1594241	1	203	11.6	8.188	1275392	318849
bho	1594958	1	66	15.327	4.532	1275966	318992
bn	1594943	1	173	11.667	2.665	1275954	318989
doi	1594902	1	294	14.583	6.388	1275921	318981
gom	1594962	1	283	11.688	20.532	1275969	318993
gu	1595249	1	294	13.15	2.299	1276199	319050
hi	1594971	1	104	14.593	2.639	1275976	318995
kn	1595303	1	95	9.868	3.091	1276242	319061
lus	1594861	1	244	15.426	78.056*	1275888	318973
mai	1594903	1	281	14.38	5.189	1275922	318981
ml	1594831	1	292	9.26	3.135	1275864	318967
mni	1594866	1	285	12.005	12.906	1275892	318974
mr	1594835	1	253	11.211	1.466	1275868	318967
ne	1595206	1	265	11.256	4.668	1276164	319042
or	1594923	1	299	11.466	5.183	1275938	318985
pa	1594855	1	180	14.55	2.33	1275884	318971
sa	1594884	1	316	11.466	22.606	1275907	318977
sd	1595391	1	291	14.645	2.01	1276312	319079
si	1594793	1	287	10.852	2.849	1275834	318959
ta	1594830	1	149	10.055	3.268	1275864	318966
te	1594724	1	72	9.618	3.544	1275779	318945
ur	1595401	1	297	15.777	1.37	1276320	319081

Table 2: Statistics of the translated dataset. In the table, the columns **Lang**, **#sample**, **#min**, **#max**, **#avg**, **%eng**, **#train**, and **#test** correspond to various attributes of the translated dataset. Specifically, they represent the language of the translated dataset, the total number of text samples, the minimum word count, the maximum word count, and the average word count in a particular dataset, the percentage of English words remaining in the translated dataset, the sample count in the train split, and the sample count in the test split, respectively.

* The percentage of English word is relatively high as Mizo (lus) language is also using Roman script. For all other languages we have used the script based approach to distinguish between English and non-English word and in case of Mizo simple dictionary based approach is applied.

for sentiment analysis in Bengali. These data were gathered from diverse sources like Bengali news articles, TV channel news dumps, books, blogs, sports portals, and manually annotated. SentNoB (Islam et al., 2021) introduced another dataset of 15K instances for analyzing sentiment in Bangla text. This dataset was created by annotating comments on news and videos from different domains.

A dataset, consisting of 15K manually annotated comments for sentiment analysis in Sinhala, is presented by Senevirathne et al. (2020).

More recently, indicSentiment (Doddapaneni et al., 2023) has been introduced for sentiment analysis in 13 Indian languages, including Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu. This dataset comprises 1K reviews for each language, which were manually translated and annotated from English reviews collected from various e-commerce websites.

3 Proposed Dataset

We have considered Sentiment140 (Go et al., 2009), a publicly available large dataset for sentiment analysis as source dataset. This dataset contains 1.6 million tweets in English extracted using the Twitter API. Each of these tweets has been labeled either with 0 or 4 which represents negative and positive sentiment respectively. As this dataset was collected from Twitter, a microblogging-based social media platform, the text predominantly employs informal and casual language, incorporating slang, fancy acronyms, emojis, URLs, as well as hashtags and keywords. Before translating this dataset to Indian languages URLs, emojis, and special symbols like # and @ have been removed. We have considered 22 Indian languages i.e. Assamese (as), Bengali (bn), Bhojpuri (bho), Dogri(doi), Gujarati (gu), Hindi (hi), Kannada (kn), Konkani (gom), Maithili (mai), Malayalam (ml), Marathi(Mr), Meiteilon

(Manipuri) (mni), Mizo (lus), Nepali (ne), Odia (or), Punjabi(pa), Sanskrit (sa), Sindhi (sd), Sinhala (si), Tamil (ta), Telugu (te), and Urdu (ur) for our study. We have created the dataset for each language listed above after translating the Sentiment140 dataset to corresponding language. The same sentiment labels are kept back after translating it to different Indian languages.

There are many machine translators (and or APIs) available like Google Translate, DeepL, Microsoft Translator, OpenAI's Machine Translation, and many more. We have chosen Google Translate as it supports 1000+ languages (Bapna et al., 2022) out of which 22 are above mentioned Indian languages.

To understand the level of correctness of the translated text, we also evaluate the performance of Google Translate taking different publicly available parallel corpora in Indian languages such as Facebook's Flores-200 dataset (Costa-jussà et al., 2022) that contains parallel corpora of 24 Indian languages out of which 19 languages are found to be relevant for our study, PMIndia dataset (Haddow and Kirefu, 2020) which we have found 12 languages are relevant for our study and IndicSentiment dataset (Doddapaneni et al., 2023). We have chosen 1000 random samples from each of these parallel corpora for evaluating the performance of Google Translate. We have taken the samples in English as the source sample and translated them to different Indian languages available in the particular parallel corpus. After that the BLEU (BiLingual Evaluation Understudy) score has been calculated taking the 4-Gram Overlap and the same is tabulated in Table 1. From this table, it can be observed that the low-resource Indian languages like Assamese (as), Bhjpuri (bho), Dogri (doi), Mizo (lus), Maithali(mai), Manipuri (mni) Sanskrit (sa) and Sinhala (si) are having relativly low BLEU score over other Indian languages. The statistics of the translated dataset for each language is tabulated in Table 2.

4 Model Training on Translated Data

In order to evaluate practicality of using the translated dataset for sentiment analysis, we implement two basic machine-learning models for each language – a simple classifier employing Multi-Layer Perceptron (MLP) and a sequential model-based classifier employing LSTM (Long Short-Term Memory). We have considered the simplest models, as we are interested in investigating usability of the datasets, rather than the model design.

The MLP model is composed of an input layer with dimensions matching the word embeddings, a single hidden layer with 256 nodes, followed by a dropout layer, and concluding with an output layer having 2 nodes for sentiment classification. Similarly, the LSTM-based model is tailored to process sequences, each with a length of 128 tokens. The architecture includes two stacked LSTM layers, each with a hidden size of 64. Subsequently, the model's output is concluded with an output layer having 2 nodes for sentiment classification. The choice of the LSTM model serves the purpose of exploring whether the translated datasets effectively retain sequential information.

To obtain text embeddings, we have used two distinct pre-trained word embedding models: (i) IndicBERT v2 (Doddapaneni et al., 2023), which was trained using the BERT architecture, and (ii) IndiSocialFT (Kumar et al., 2023)– a multilingual fastText (Bojanowski et al., 2017) based embeddings for Indian languges. To obtain the embedding for a specific text(tweet), we have calculated the mean of all the embeddings generated by the pre-trained model for each token. The text embedding generated by the IndicBERT v2 is of size 768 and that of IndiSocialFT is 300. These text embeddings are directly supplied to the input layer of our MLP model.

For each language in our dataset, we have trained separate classifiers, taking into account the embeddings from both IndicBERT v2 and IndiSocialFT individually (for LSTM model only IndicBERT v2 is used). We split each language dataset into an 80% portion for model training and a 20% portion for model testing. The statistics for the train and test splits for each language dataset are summarized in Table 2. During the model training process, we have employed the early-stopping technique to mitigate the risk of overfitting the data, ensuring our model's generalizability and performance.

5 Result and Evalution

We have evaluated the trained model considering two datasets - (i) the test split of the translated dataset mentioned in Table 2, and (ii) IndicSentimet dataset–a human translated and annotated dataset (Doddapaneni et al., 2023).

	IndiSocialFT					indicBert v2			
Lang	**Acc**	**Preci**	**Recall**	**F1**	**Acc***	**Acc**	**Preci**	**Recall**	**F1**
en	**0.78**	**0.78**	**0.78**	**0.78**	NA	**0.813**	**0.813**	**0.813**	**0.813**
as	0.751	0.752	0.751	0.751	0.752	0.789	0.789	0.789	0.789
bho	0.766	0.766	0.766	0.766	0.766	0.787	0.787	0.787	0.787
bn	0.769	0.769	0.769	0.769	0.773	0.798	0.798	0.798	0.798
doi	0.758	0.758	0.758	0.758	0.758	0.784	0.784	0.784	0.784
gom	0.748	0.748	0.748	0.748	0.735	0.778	0.778	0.778	0.778
gu	0.769	0.77	0.769	0.769	0.773	0.803	0.803	0.803	0.803
hi	0.76	0.76	0.76	0.76	0.763	0.792	0.792	0.792	0.792
kn	0.761	0.761	0.762	0.761	0.767	0.800	0.801	0.800	0.800
lus	0.751	0.751	0.751	0.751	–	0.747	0.747	0.747	0.747
mai	0.756	0.756	0.756	0.756	0.760	0.786	0.786	0.786	0.786
ml	0.773	0.774	0.773	0.773	0.777	0.802	0.802	0.802	0.802
mni	0.734	0.734	0.733	0.733	0.736	0.718	0.717	0.718	0.718
mr	0.76	0.76	0.76	0.76	0.762	0.796	0.796	0.796	0.796
ne	0.725	0.725	0.725	0.725	0.733	0.774	0.774	0.774	0.774
or	0.749	0.75	0.749	0.749	0.750	0.779	0.779	0.779	0.779
pa	0.755	0.755	0.755	0.755	0.756	0.795	0.795	0.795	0.795
sa	0.75	0.75	0.75	0.75	0.735	0.778	0.778	0.778	0.778
sd	0.728	0.728	0.728	0.727	0.730	0.772	0.772	0.772	0.772
si	0.722	0.722	0.722	0.722	0.729	0.700	0.700	0.700	0.700
ta	0.763	0.764	0.763	0.763	0.765	0.792	0.792	0.792	0.792
te	0.766	0.766	0.766	0.766	0.770	0.792	0.792	0.792	0.791
ur	0.745	0.745	0.745	0.745	0.748	0.792	0.792	0.792	0.792
avg	0.753	0.753	0.753	0.752	–	0.780	0.780	0.780	0.780
	bert-base-uncased								
en	**0.822**	**0.822**	**0.822**	**0.822**					

Table 3: Sentiment analysis over the dataset obtained after translating the Sentiment140 dataset using Google Translate. The performance metrics reported as accuracy (**Acc**), micro precision (**Preci**), micro recall (**Recall**) and micro F1 (**F1**) of the model trained using the **IndiSocialFT**, and **indicBert v2**. (**Acc*** is accuracy after removing the untranslated English word).

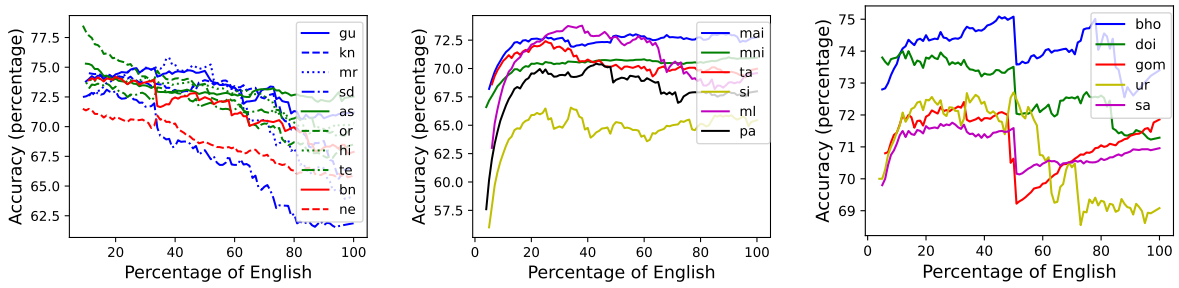
5.1 Evaluation with Translated Sentimet140

5.1.1 Evaluation with MLP model

We have conducted an evaluation on 20% of the entire Google-translated Sentiment140 dataset for each language to assess the performance of the trained MLP model for sentiment analysis. Similar to the model training phase, we have utilized both fastText and IndicBERT v2 embedding models independently for embedding the test set. The evaluation results, encompassing various performance metrics such as accuracy, micro precision, micro recall, and micro F1, are presented in Table 3.

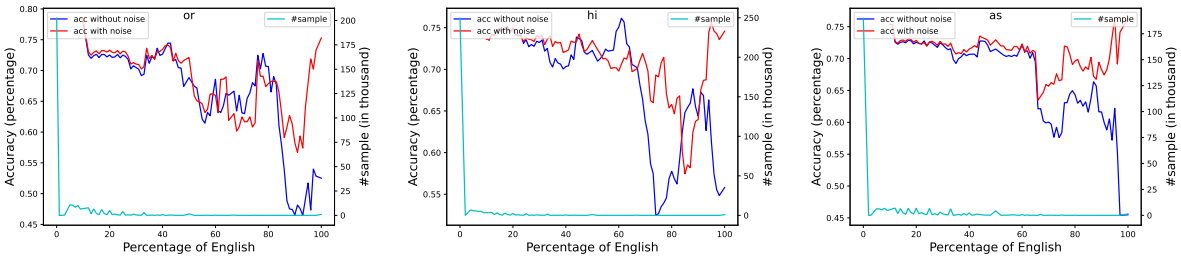
From the table, it is apparent that when evaluating the fastText-based MLP model, accuracy remains consistently high across all languages, approximately at $75\% \pm 2\%$. Similarly, other performance metrics also exhibit remarkable consistency. When assessing the indicBert-based MLP model, we find a consistent accuracy of $76\% \pm 5\%$ across all languages. This remarkable consistent performance suggests that, despite the lower BLEU scores for low-resource languages, Google Translate effectively preserves the sentiment in the translated text.

Notably, in both setups, the model's performance on the English-language dataset, i.e., the original Sentiment140 dataset, stands out high. We have also observed that IndicBert v2 performs better than the fastText one. We have got the correlation value of 0.738 between the accuracy of model trained over fastText embedding and that of IndicBert v2 embedding. This high correlation value



(a)Negative Slope (b)Positive Slope (c)No consistent pattern

Figure 1: Graph of Performance of IndiSocialFT based model in terms of Accuracy vs Percentage of English left in the translated Sentiment140 dataset for various languages.



(a)or (b)hi (c)as

Figure 2: Graph of Accuracy with noise and without noise over the same sample of IndiSocialFT-based model for Odia (or), Hindi (hi), and Assamese (as) language.

	IndicBert v2			
Lang	**Acc**	**Preci**	**Recall**	**F1**
en	**0.83**	**0.83**	**0.83**	**0.83**
as	0.807	0.808	0.807	0.807
bho	0.811	0.812	0.811	0.811
bn	0.810	0.812	0.810	0.810
doi	0.812	0.812	0.812	0.812
gom	0.8	0.802	0.8	0.799
gu	0.816	0.816	0.816	0.816
hi	0.813	0.813	0.813	0.813
kn	0.814	0.814	0.814	0.814
lus	0.802	0.802	0.802	0.802
mai	0.807	0.807	0.807	0.807
ml	0.813	0.813	0.813	0.813
mni	0.788	0.787	0.787	0.787
mr	0.807	0.807	0.807	0.807
ne	0.788	0.788	0.788	0.788
or	0.794	0.794	0.794	0.794
pa	0.806	0.807	0.806	0.805
sa	0.801	0.801	0.801	0.801
ta	0.804	0.804	0.804	0.804
te	0.806	0.806	0.806	0.806
ur	0.811	0.811	0.811	0.811

Table 4: Evaluation of trained model on Translated Sentiment140 using LSTM based model and IndicBert v2. The performance metrics reported are accuracy (**Acc**), micro precision (**Preci**), micro recall (**Recall**) and micro F1 (**F1**).

indicates that both the models perform in same way for all the languages. Only the difference is that they have different embedding space. This difference in performance with different embeddings suggest that the context of sentiment analysis is closely intertwined with the choice of embedding space.

We have also calculated the correlation between the accuracy of both model and the percentage of English remaining in each languages. We have got the correlation value of -0.183 and -0.243 while considering fastText and indicBert v2 embedding respectively. This negative value of the correlation suggests that there is no relation between the Accuracy and the percentage of the English present in the translated text.

For the purpose of comparison, we have also trained the basic MLP model using the original Sentiment140 dataset with three different pre-trained embeddings: fastText, indicBert v2, and bert-base-uncased (Devlin et al., 2018) and the performance of these models is tabulated in Table 3. The performance of these models is found to be similar, with bert-base-uncased exhibiting slightly better results.

For error analysis, we have also generated a graphical representation illustrating the perfor-

		Human Translated				Google Translated					
Lang	**#sample**	**#min**	**#max**	**#avg**	**%eng**	**BLEU**	**#min**	**#max**	**#avg**	**%eng**	**corBE**
as	1000	2	80	20.281	3.538	14.306	2	77	19.845	5.284	-0.034
bn	1000	2	88	20.67	0	29.445	2	282	20.509	2.631	-0.043
gu	1000	2	93	22.997	1.725	35.926	2	92	21.977	3.788	-0.01
hi	1000	3	100	26.756	0.423	34.001	2	107	26.833	0.421	-0.014
kn	1000	2	67	16.974	0.134	29.845	2	58	16.138	4.284	-0.064
ml	1000	2	62	15.882	3.59	21.961	2	53	14.625	5.12	-0.04
mr	1000	2	74	20.081	2.058	7.746	2	79	18.879	1.736	0.081
or	1000	2	86	20.649	5.206	30.084	2	69	20.132	11.311	-0.052
pa	1000	2	106	26.498	1.646	34.531	2	82	25.395	2.519	-0.012
ta	1000	2	69	17.473	4.06	23.922	2	71	17.054	4.071	0.078
te	1000	2	90	18.415	3.048	17.092	2	74	17.197	4.635	0.015
ur	1000	3	118	29.119	0.2	39.102	2	121	28.781	1.885	0.004

Table 5: Statistics of IndicSentiment Dataset. In the table, the columns **Lang**, **#sample**, **#min**, **#max**, **#avg**, and **%eng** correspond to various attributes of the translated dataset. Specifically, they represent the language, the total number of text samples, the minimum word count, the maximum word count, the average word count, and the percentage of English words remaining in the translated dataset respectively. The column **BLEU** reports the BLEU score of the Google translated dataset and the provided human translated dataset. The column **corBE** reports the correlations of the BLEU score and the percentage of English words remaining in the Google translated dataset.

mance of the IndiSocialFT-based model in terms of accuracy concerning the percentage of remaining English words in the translated Sentiment140 dataset for various languages. This analysis allows us to delve deeper into the structure of the translated dataset. The corresponding graph is displayed in Fig. 1.

In the Graph presented in Fig. 1.a, we observe a collection of 10 languages, each exhibiting an average negative slope. This pattern indicates that as the percentage of untranslated words increases, the model's performance declines. In contrast, the Graph in Fig. 1.b shows a set of 6 languages where the performance initially improves and then stabilizes as the percentage of untranslated words increases. Finally, in Fig. 1.c, we find the plot of 5 languages for which the relationship between performance and the percentage of untranslated words does not adhere to a consistent pattern. Analyzing these graphs, we can draw a general conclusion that most translated datasets exhibit a common behavior: as the level of noise in the dataset increases, performance tends to decrease. This decline in performance can be attributed to the fact that the embedding space of the noise (in this case, untranslated English words) differs significantly from the embedding space of the words in a particular language.

Further, we have also plotted the graph of accuracy with noise and accuracy after removing the untranslated English word(i.e. noise) over the same sample for some of the languages and the results are shown in Fig. 2. These plots illustrate that the variation of the accuracy remains consistent for most of the sample. Most of the variation is observed for the sample having greater than 70% of the noise and these samples are very few in number. This observation also suggests that this translated dataset has the potential to serve as a foundational framework for tackling the low-resource challenges commonly encountered in sentiment analysis for Indian languages.

5.1.2 Evaluation with LSTM model

We have conducted an evaluation on 20% of the entire Google-translated Sentiment140 dataset for each language to assess the performance of the trained LSTM model for sentiment analysis. while training and testing, we have only utilized the IndicBERT v2 embedding model for text embedding. The results of the evaluation, including various performance metrics such as accuracy, micro precision, micro recall, and micro F1, are presented in Table 4.

Upon comparing the indicBert v2 column of Table 3 with Table 4, we observe a notable improvement in results following the utilization of the sequential model, i.e., LSTM. This significant performance enhancement suggests that Google Translate adequately preserves the sequence in the translated text across all languages.

	Metrics \ lang	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	ur
IndiSocialFT	**Acc**	0.728	0.770	0.763	0.763	0.710	0.763	0.746	0.747	0.757	0.772	0.777	0.760
	Preci	0.731	0.774	0.763	0.764	0.713	0.766	0.749	0.747	0.757	0.773	0.778	0.760
	Recall	0.727	0.769	0.763	0.763	0.709	0.762	0.745	0.747	0.757	0.772	0.777	0.760
	F1	0.727	0.769	0.763	0.763	0.708	0.762	0.745	0.747	0.757	0.772	0.777	0.760
indicBert v2	**Acc**	0.879	0.873	0.893	0.897	0.886	0.897	0.879	0.858	0.888	0.889	0.894	0.886
	Preci	0.879	0.881	0.894	0.898	0.887	0.900	0.880	0.859	0.888	0.890	0.895	0.886
	Recall	0.879	0.872	0.893	0.897	0.886	0.896	0.879	0.858	0.888	0.889	0.894	0.886
	F1	0.879	0.872	0.893	0.897	0.886	0.897	0.879	0.858	0.888	0.889	0.894	0.886

Table 6: Evaluation of the trained model on the IndicSentiment dataset. The performance metrics accuracy (**Acc**), micro precision (**Preci**), micro recall (**Recall**), and micro F1 (**F1**) of the model trained using IndiSocialFT are reported under **IndiSocialFT**, and those using indicBert v2 are under **indicBert v2**.

	Metrics \ lang	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	ur
IndiSocialFT	**Acc**	0.728	0.753	0.745	0.764	0.703	0.750	0.736	0.746	0.741	0.752	0.763	0.733
	Preci	0.729	0.756	0.746	0.764	0.707	0.755	0.738	0.747	0.743	0.752	0.764	0.733
	Recall	0.727	0.752	0.744	0.764	0.702	0.749	0.735	0.745	0.740	0.752	0.762	0.733
	F1	0.727	0.752	0.744	0.764	0.701	0.748	0.735	0.745	0.740	0.752	0.762	0.733
indicBert v2	**Acc**	0.853	0.863	0.883	0.911	0.859	0.847	0.877	0.853	0.858	0.863	0.863	0.885
	Preci	0.853	0.868	0.885	0.911	0.859	0.852	0.878	0.853	0.859	0.864	0.865	0.886
	Recall	0.853	0.862	0.883	0.911	0.859	0.846	0.877	0.853	0.858	0.863	0.862	0.885
	F1	0.853	0.862	0.883	0.911	0.859	0.846	0.877	0.853	0.858	0.863	0.862	0.885

Table 7: Evaluation of the trained sentiment analysis model over the dataset obtained after translating the IndicSentiment dataset using Google Translate considering text in English as the source text. The performance metrics accuracy (**Acc**), micro precision (**Preci**), micro recall (**Recall**), and micro F1 (**F1**) of the model trained using IndiSocialFT are reported under **IndiSocialFT**, and those using indicBert v2 are under **indicBert v2**.

5.2 Performance with IndicSentiment Dataset

Taking into account the socio-cultural and language-specific intricacies that impact sentiment, we go beyond merely employing the translated corpus. We also integrate a human-translated sentiment dataset, indicSentiment (Doddapaneni et al., 2023), into our evaluation of the sentiment analysis model trained on the translated corpus, *IndiSentiment140*. IndicSentiment dataset consists of the text available in 13 Indian languages which have been manually translated to these different languages from same English source text (Doddapaneni et al., 2023). Out of these 13 languages only 12 languages dataset are found to be relevant for our study. For each languages we have considered the 1000 sample for the evaluation purpose. Each of these samples is having either positive or negative sentiment label.

While considering this IndicSentiment dataset for evaluating the trained MLP model we have considered two setup. In first setup we have considered the provided human translated text as testing dataset and in another setup we have used the Google Translate to translate the provided English text to the different Indian languages available in this parallel corpus. Then this Google translated version of IndicSentiment is also used for further evaluation. A detailed statistics of IndicSentimet dataset both the Human Translated version and the Google Translated version is tabulated in Table 5. In the Table we have also mentioned the correlations of the BLEU score and the percentage of English word remaining in the translated text. Most of these correlation values are either negative or near to zero that suggests there is no link of BLEU score with the percentage of English word remains in the translated dataset.

The performance of the trained MLP sentiment classifier model in term of different performance metrics i.e. accuracy, micro precision, micro recall and micro F1 considering both pre-trained embedding models over the manually (human) translated version of IndicSentiment is reported in Table 6 and that of Google translated version is in Table 7. In these tables also all performance metrics remain consistent.

Readings from the Table 6 suggest that the model trained over the translated dataset for each language is efficient enough for predicting the sentiment for the manually translated (naturally written) data. Here also we find the same trend for both MLP model i.e. based on fastText and the indicBert v2, across all languages. The accuracy of the fastText based model remains consistent, $74\% \pm 3\%$ across all languages, and a consistent accuracy of $87\% \pm 2\%$ across all languages for indicBert v2 based MLP model.

Results in Table 7 also follow the same trend that again suggests that the machine translation (Google Translate) preserves the sentiment of the text.

6 Conclusion and Future Work

In this paper, we have investigated the potential of machine translation as a solution to overcome the low-resource challenge faced by diverse Indian languages, specifically in the context of sentiment analysis. Our approach involved translating the Sentiment-140 dataset into 22 Indian languages using Google Translate and conducting comprehensive experiments to assess the usability of these translated datasets. Our findings reveal that these translated datasets not only retain the sentiment characteristics of the original dataset but also maintain the sequential structure within the translated text across specific languages. This suggests that these translated datasets could serve as a foundational framework to address the low-resource issue prevalent in Indian languages.

As for future work, we plan to further improve our dataset by incorporating more machine translation tools and APIs available for Indian languages. We also plan to add more languages to this dataset.

Limitations

While the translated dataset proposed in this study serves its intended purpose, it is essential to acknowledge certain limitations. One notable limitation is the presence of noise in the form of untranslated English words within the dataset. This factor may limit its applicability for diverse linguistic purposes. Moreover, conducting a direct comparison between the proposed translated dataset and a manually translated dataset—representing the actual translation in the target language—poses a challenge. Unfortunately, as of now, there is no available human-translated dataset for this specific content. Consequently, the absence of such a benchmark hinders a comprehensive assessment of the accuracy and quality of the automated translation.

Acknowledgements

This study is partially funded by Ministry of Electronics and Information Technology, Government of India, New Delhi, India.

References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in hindi: resource creation and evaluation. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 2703–2709.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022. Building machine translation systems for the next thousand languages. Technical report, Google Research.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Rama Rohit Reddy Gangula and Radhika Mamidi. 2018. Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Saurabh Kumar, Ranbir Sanasam, and Sukumar Nandi. 2023. IndiSocialFT: Multilingual word representation for Indian languages in code-mixed environment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3866–3871, Singapore. Association for Computational Linguistics.

Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, and Rajiv Ratn Shah. 2019. Bhaav-a text corpus for emotion analysis from hindi stories. *arXiv preprint arXiv:1910.04073*.

John P. McCrae Md. Rezaul Karim, Bharathi Raja Chakravarti and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network.

Lahiru Senevirathne, Piyumal Demotte, Binod Karunanayake, Udyogi Munasinghe, and Surangika Ranathunga. 2020. Sentiment analysis for sinhala language using deep learning techniques. *arXiv preprint arXiv:2011.07280*.