AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages

Anoop Kunchukuttan[1], **Divyanshu Kakwani**[2], **Satish Golla**[3], **Gokul N.C.**[4],
Avik Bhattacharyya[5], **Mitesh M. Khapra**[6], **Pratyush Kumar**[7]
Microsoft India[1], IIT Madras[2,6,7], AI4Bharat[3,4,5]
ankunchu@microsoft.com[1], gokulnc@ai4bharat.org[4],
{gsatishkumaryadav,avikbhattacharyya.2k}@gmail.com[3,5],
{divk,miteshk,pratyush}@cse.iitm.ac.in[2,6,7]

Abstract

We present the IndicNLP corpus, a large-scale, general-domain corpus containing 2.7 billion words for 10 Indian languages from two language families. We share pre-trained word embeddings trained on these corpora. We create news article category classification datasets for 9 languages to evaluate the embeddings. We show that the IndicNLP embeddings significantly outperform publicly available pre-trained embedding on multiple evaluation tasks. We hope that the availability of the corpus will accelerate Indic NLP research. The resources are available at https://github.com/ai4bharat-indicnlp/indicnlp_corpus.

1 Introduction

Distributional representations are the corner stone of modern NLP, which have led to significant advances in many NLP tasks like text classification, NER, sentiment analysis, MT, QA, NLI, *etc.* Particularly, word embeddings (Mikolov et al., 2013b), contextualized word embeddings (Peters et al., 2018), and language models (Devlin et al., 2019) can model syntactic/semantic relations between words and reduce feature engineering. These pre-trained models are useful for initialization and/or transfer learning for NLP tasks. They are useful for learning multilingual embeddings, enabling cross-lingual transfer. Pre-trained models are typically learned using unsupervised approaches from large, diverse monolingual corpora. The quality of embeddings is impacted by the size of the monolingual corpora (Mikolov et al., 2013a; Bojanowski et al., 2017), a resource not widely available publicly for many major languages.

Indic languages, widely spoken by more than a billion speakers, lack large, publicly available monolingual corpora. They include 8 out of top 20 most spoken languages and ∼30 languages with more than a million speakers. There is also a growing population of users consuming Indian language content (print, digital, government and businesses).

Indic languages are very diverse, spanning 4 major language families. The Indo-Aryan and Dravidian languages are spoken by 96% of the population in India. The other families are diverse, but the speaker population is relatively small. Almost all languages have SOV word order and are morphologically rich. The language families have also interacted over a long period of time leading to significant convergence in linguistic features; hence, the Indian subcontinent is referred to as a *linguistic area* (Emeneau, 1956). So, Indic languages are of great interest and importance for NLP research.

In this work, we address the creation of large, general-domain monolingual corpora for multiple Indian languages. Using the new monolingual corpora, we also create other essential resources for Indian language NLP. We evaluate these resources on various benchmarks, in the process creating some new evaluation benchmarks. This work contributes the following Indic language resources:

- A large monolingual corpora (IndicNLP corpus) for 10 languages from two language families (Indo-Aryan branch and Dravidian). Each language has at least 100 million words (except Oriya).
- Pre-trained word embeddings for 10 Indic languages trained using FastText.
- News article category classification datase for 9 languages.
- Unsupervised morphanalyzers for 10 languages.

We show that IndicNLP embeddings outperform publicly available embeddings on various tasks: word similarity, word analogy, sentiment analysis, text classification, bilingual lexicon induction. Further, we show the utility of the monolingual corpora for training morphanalzyers.

Lang	#Sentences	#Tokens	#Types
pa	6,525,312	179,434,326	594,546
hi	62,961,411	1,199,882,012	5,322,594
bn	7,209,184	100,126,473	1,590,761
or	3,593,096	51,592,831	735,746
gu	7,889,815	129,780,649	2,439,608
mr	9,934,816	142,415,964	2,676,939
kn	14,780,315	174,973,508	3,064,805
te	15,152,277	190,283,400	4,189,798
ml	11,665,803	167,411,472	8,830,430
ta	20,966,284	362,880,008	9,452,638
Total	160,678,313	2,698,780,643	38,897,865

Table 1: IndicNLP corpus statistics

2 Related Work

Text Corpora. Few organized sources of monolingual corpora exist for most Indian languages. The EMILLE/CIIL corpus (McEnery et al., 2000) was an early effort to build a corpora for South Asian languages, spanning 14 languages with a total of 92 million words. *Wikipedia* for Indian languages is small (the largest one, Hindi, has just 40 million words). The Leipzig corpus (Goldhahn et al., 2012) contains small collections of upto 1 million sentences for news and web crawls (average 300K sentences). In addition, there are some language specific corpora for Hindi and Urdu (Bojar et al., 2014; Jawaid et al., 2014).

The *CommonCrawl* project crawls webpages in many languages by sampling various websites. Our analysis of a processed crawl for the years 2013-2016 (Buck et al., 2014) for Indian languages revealed that most Indian languages, with the exception of Hindi, Tamil and Malayalam, have few good sentences ($\geq$10 words) - in the order of around 50 million words. The Indic corpora size has not possibly changed siginificantly in recent years based on the recent CommonCrawl statistics.

Word Embeddings. Word embeddings have been trained for many Indian languages using limited corpora. The Polyglot (Al-Rfou et al., 2013) and FastText projects provide embeddings trained on Wikipedia. FastText also provides embeddings trained on Wikipedia + CommonCrawl corpus.

3 IndicNLP Corpus Creation

We describe the creation of the *IndicNLP* corpus.

Data sources. Our goal is collection of corpora that reflects contemporary use of Indic languages and covers a wide range of topics. Hence, we focus primarily on the news domain and Wikipedia. We source our data from popular Indian languages news websites, identifying many news sources from *W3Newspapers*[1]. We augmented our crawls with some data from other sources: Leipzig corpus (Goldhahn et al., 2012) (Tamil and Bengali), WMT NewsCrawl (for Hindi), WMT CommonCrawl (Buck et al., 2014) (Tamil, Malayalam), HindEnCorp (Hindi) (Bojar et al., 2014).

News Article Crawling. We use *Scrapy*[2], a web-crawling Python framework, for crawling news websites. If the site has a good sitemap, we rely on it to increase the efficiency of the crawler. For other sources, we crawl all the links recursively.

Article Extraction. For many news websites, we used *BoilerPipe*[3], a tool to automatically extract the main article content for structured pages without any site-specific learning and customizations (Kohlschütter et al., 2010). This approach works well for most of the Indian language news websites. In some cases, we wrote custom extractors for each website using *BeautifulSoup*[4], a Python library for parsing HTML/XML documents. After content extraction, we applied some filters on content length, script *etc.*, to select good quality articles. We used the *wikiextractor*[5] tool for text extraction from Wikipedia.

Text Processing. First, we canonicalize the representation of Indic language text in order to handle multiple Unicode representations of certain characters and typing inconsistencies. Next, we sentence split the article and tokenize the sentences. These steps take into account Indic punctuations and sentence delimiters. Heuristics avoid creating sentences for initials (P. G. Wodehouse) and common Indian titles (Shri., equivalent to Mr. in English) which are followed by a period. We use the *Indic NLP Library*[6] (Kunchukuttan, 2020) for processing.

The final corpus collection is created after deduplicating and shuffling sentences. We used the Murmurhash algorithm (*mmh3* python library with a 128-bit unsigned hash) for de-duplication.

Dataset Statistics. Table 1 shows statistics of the

[1] https://www.w3newspapers.com
[2] https://scrapy.org/
[3] https://github.com/kohlschutter/boilerpipe
[4] https://www.crummy.com/software/BeautifulSoup
[5] https://github.com/attardi/wikiextractor
[6] https://github.com/anoopkunchukuttan/indic_nlp_library

monolingual datasets for each language. Hindi and Tamil are the largest collections, while Oriya has the smallest collection. All other languages have a collection between 100-200 million words. Bengali, a widely spoken language, has only around 100 million words: we would like to increase that collection. The Hindi corpus is a compilation of existing sources. CommonCrawl is a significant contributor to the Tamil corpus (55%) and Malayalam (35%). Most of the data for other languages originate from our crawls.

4 IndicNLP Word Embeddings

We train pre-trained word embeddings using the IndicNLP corpus, and evaluate their quality on: (a) word similarity, (b) word analogy, (c) text classification, (d) bilingual lexicon induction tasks. We compare the IndicNLP embeddings with two pre-trained embeddings released by the *FastText* project trained on Wikipedia (*FT-W*) (Bojanowski et al., 2017) and Wiki+CommonCrawl (*FT-WC*) (Grave et al., 2018) respectively. This section describes the training of word embeddings, and evaluation settings and results for each task.

4.1 Training Details

We train 300-dimensional word embeddings for each language on the IndicNLP corpora using *FastText* (Bojanowski et al., 2017). Since Indian languages are morphologically rich, we chose *FastText*, which is capable of integrating subword information by using character n-gram embeddings during training.

We train skipgram models for 10 epochs with a window size of 5, minimum token count of 5 and 10 negative examples sampled for each instance. We chose these hyper-parameters based on suggestions by Grave et al. (2018). Otherwise, default settings were used. Based on previously published results, we expect FastText to be better than word-level algorithms like *word2vec* (Mikolov et al., 2013b) and *GloVe* (Pennington et al., 2014) for morphologically rich languages. We leave comparison with word-level algorithms for future work.

4.2 Word Similarity & Analogy Evaluation

We perform an intrinsic evaluation of the word embeddings using the IIIT-Hyderabad word similarity dataset (Akhtar et al., 2017) which contains similarity databases for 7 Indian languages. The database contains similarity judgments for around 100-200 word-pairs per language. Table 2 shows the evaluation results. We also evaluated the Hindi word embeddings on the Facebook Hindi word analogy dataset (Grave et al., 2018). IndicNLP embeddings outperform the baseline embeddings on an average.

Lang	FT-W	FT-WC	INLP
Word Similarity (*Pearson Correlation*)			
pa	**0.467**	0.384	0.428
hi	0.575	0.551	**0.626**
gu	0.507	0.521	**0.614**
mr	0.497	**0.544**	0.495
te	0.559	0.543	**0.560**
ta	**0.439**	0.438	0.392
Average	0.507	0.497	**0.519**
Word Analogy (*% accuracy*)			
hi	19.76	32.93	**33.48**

Table 2: Word Similarity and Analogy Results

Lang	Classes	# Articles	
		Train	Test
pa	BIZ, ENT. POL, SPT	2,496	312
bn	ENT, SPT	11,200	1,400
or	BIZ, CRM, ENT, SPT	24,000	3,000
gu	BIZ, ENT, SPT	1,632	204
mr	ENT, STY, SPT	3,815	478
kn	ENT, STY, SPT	24,000	3,000
te	ENT, BIZ, SPT	19,200	2,400
ml	BIZ, ENT, SPT, TECH	4,800	600
ta	ENT, POL, SPT	9,360	1,170

Table 3: IndicNLP News category dataset statistics. The following are the categories: entertainment: ENT, sports: SPT, business: BIZ, lifestyle; STY, techology: TECH, politics: POL, crime: CRM

4.3 Text Classification Evaluation

We evaluated the embeddings on different text classification tasks: (a) news article topic, (b) news headlines topic, (c) sentiment classification. We experimented on publicly available datasets and a new dataset (IndicNLP News Category dataset).

Publicly available datasets. We used the following datasets: (a) IIT-Patna Sentiment Analysis dataset (Akhtar et al., 2016), (b) ACTSA Sentiment Analysis corpus (Mukku and Mamidi, 2017), (c) BBC News Articles classification dataset, (d) iNLTK Headlines dataset, (e) Soham Bengali News

Lang	FT-W	FT-WC	INLP
pa	94.23	94.87	**96.79**
bn	97.00	97.07	**97.86**
or	94.00	95.93	**98.07**
gu	97.05	97.54	**99.02**
mr	96.44	97.07	**99.37**
kn	96.13	96.50	**97.20**
te	98.46	98.17	**98.79**
ml	90.00	89.33	**92.50**
ta	95.98	95.81	**97.01**
Average	95.47	95.81	**97.40**

Table 4: Accuracy on IndicNLP News category testset

Lang	Dataset	FT-W	FT-WC	INLP
hi	BBC Articles	72.29	67.44	**74.25**
	IITP+ Movie	41.61	44.52	**45.81**
	IITP Product	58.32	57.17	**63.48**
bn	Soham Articles	62.79	64.78	**72.50**
gu	iNLTK Headlines	81.94	84.07	**90.90**
ml		86.35	83.65	**93.49**
mr		83.06	81.65	**89.92**
ta		90.88	89.09	**93.57**
te	ACTSA	46.03	42.51	**48.61**
	Average	69.25	68.32	**74.73**

Table 5: Text classification accuracy on public datasets

	en to Indic			**Indic to en**		
	FT-W	**FT-WC**	**INLP**	**FT-W**	**FT-WC**	**INLP**
bn	22.60	**33.92**	33.73	31.22	**42.10**	41.90
hi	40.93	44.35	**48.69**	49.56	57.16	**58.93**
te	21.10	23.01	**29.33**	25.36	32.84	**36.54**
ta	19.27	30.25	**34.43**	26.66	40.20	**42.39**
Ave.	25.98	32.88	**36.55**	33.20	43.08	**44.94**

Table 6: Accuracy@1 for bilingual lexicon induction

classification dataset. Details of the datasets can be found in the Appendix A. Our train and test splits derived from the above mentioned corpora are made available on the IndicNLP corpus website.

IndicNLP News Category Dataset. We use the IndicNLP corpora to create classification datasets comprising news articles and their categories for 9 languages. The categories are determined from URL components. We chose generic categories like entertainment and sports which are likely to be consistent across websites. The datasets are balanced across classes. See Table 3 for details.

Classifier training. We use a k-NN ($k = 4$) classifier since it is a non-parameteric - the classification performance directly reflects the how well the embedding space captures text semantics (Meng et al., 2019). The input text embedding is the mean of all word embeddings (bag-of-words assumption).

Results. On nearly all datasets & languages, IndicNLP embeddings outperform baseline embeddings (See Tables 4 and 5).

4.4 Bilingual Lexicon Induction

We use IndicNLP embeddings for creating mutlilingual embeddings, where monolingual word embeddings from different languages are mapped into the same vecor space. Cross-lingual learning using multilingual embeddings is useful for Indic languages which are related and training data for NLP tasks is skewed across languages. We train bilingual word embeddings from English to Indian languages and vice versa using GeoMM (Jawanpuria et al., 2019), a state-of-the-art supervised method for learning bilingual embeddings. We evaluate the bilingual embeddings on the BLI task, using bilingual dictionaries from the MUSE project and *en-te* dictionary created in-house. We search among the 200k most frequent target language words with the CSLS distance metric during inference (Conneau et al., 2018). The quality of multilingual embeddings depends on the quality of monolingual embeddings. IndicNLP bilingual embeddings significantly outperform the baseline bilingual embeddings (except Bengali).

5 Unsupervised Morphology Analyzers

Indian languages are morphologically rich. The large vocabulary poses data sparsity problems for NLP applications. Morphological analysis provides a means to factor the words into its constituent morphemes. However, morphanalyzers are either not available for many Indic languages or have limited coverage and/or low quality. Significant linguistic expertise and effort is need to build morphanalyzers for all major Indic languages. On the other hand, unsupervised morphanalyzers can be easily built using monolingual corpora.

Training. We trained unsupervised morphanalyz-

Lang Pair	word	morph (K&B, 2016)	morph (INLP)
bn-hi	31.23	32.17	**33.80**
pa-hi	68.96	71.29	**71.77**
hi-ml	8.49	9.23	**9.41**
ml-hi	15.23	17.08	**17.62**
ml-ta	6.52	**7.61**	7.12
te-ml	6.62	**7.86**	7.68
Average	22.84	24.21	**24.57**

Table 7: SMT between Indian languages (BLEU scores)

ers using Morfessor 2.0. (Virpioja et al., 2013). We used only the word types (with minimum frequency=5) without considering their frequencies for training. This configuration is recommended when no annotated data is available for tuning. **SMT at morph-level.** We consider SMT between Indic languages as a usecase for our morphanalzyers. We compare word-level models and two morph-level models (trained on IndicNLP corpus and Kunchukuttan and Bhattacharyya (2016)'s model) to verify if our morphanalyzers can address data sparsity issues. We trained Phrase-based SMT models on the ILCI parallel corpus (Jha, 2012) (containing about 50k sentences per Indian language pair). We use the same data and training configuration as Kunchukuttan and Bhattacharyya (2016).

Results. We see that our model outperforms the word-level model significantly, while it outperforms the Kunchukuttan and Bhattacharyya (2016)'s morphanalyzer in most cases (results in Table 7). Note that the their morphanalyzer is trained on data containing the parallel corpus itself, so it may be more tuned for the task. Thus, we see that the IndicNLP corpora can be useful for building morphological analyzers which can benefit downstream tasks.

6 Summary and Future Work

We present the IndicNLP corpus, a large-scale, general-domain corpus of 2.7 billion words across 10 Indian languages, along with word embeddings, morphanalyzers and text classification benchmarks. We show that resources derived from this corpus outperform other pre-trained embeddings and corpora on many NLP tasks. The corpus, embeddings and other resources will be publicly available for research.

We are working on expanding the collection to at least 1 billion words for major Indian languages. We further plan to build: (a) richer pre-trained representations (BERT, ELMo), (b) multilingual pre-trained representations, (c) benchmark datasets for representative NLP tasks. While these tasks are work-in-progress, we hope the availability of this corpus will accelerate NLP research for Indian languages by enabling the community to build further resources and solutions for various NLP tasks and opening up interesting NLP questions.

References

Md. Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 482–493.

Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava, and Manish Shrivastava. 2017. Word similarity datasets for Indian languages: Annotation and baseline systems. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 91–94, Valencia, Spain. Association for Computational Linguistics.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Balamurali A.R., Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for Indian languages using linked WordNets. In *Proceedings of COLING 2012: Posters*, pages 73–82, Mumbai, India. The COLING 2012 Organizing Committee.

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.

Ondrej Bojar, Vojtech Diatka, Pavel Rychlỳ, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. 2014. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pages 3550–3555.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Murray B Emeneau. 1956. India as a lingustic area. *Language*.

D. Goldhahn, T. Eckart, and U. Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012)*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Bushra Jawaid, Amir Kamran, and Ondrej Bojar. 2014. A tagged corpus and a tagger for urdu. In *LREC*, pages 2938–2943.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: a geometric approach. *Transaction of the Association for Computational Linguistics (TACL)*, 7:107–120.

Girish Nath Jha. 2012. The TDIL program and the Indian Language Corpora Initiative. In *Language Resources and Evaluation Conference*.

Christian Kohlschütter, Péter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *WSDM*.

Anoop Kunchukuttan. 2020. The IndicNLP Library. Indian language NLP Library.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Orthographic Syllable as basic unit for SMT between Related Languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Anthony McEnery, Paul Baker, Rob Gaizauskas, and Hamish Cunningham. 2000. Emille: Building a corpus of south asian languages. *VIVEK-BOMBAY-*, 13(3):22–28.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. 2019. Spherical text embedding. In *Advances in Neural Information Processing Systems*, pages 8206–8215.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sandeep Sricharan Mukku and Radhika Mamidi. 2017. ACTSA: Annotated corpus for Telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Aalto University.

A Publicly Available Text Classification Datasets

We used the following publicly available datasets for our text classification experiments:

(a) IIT-Patna Movie and Product review dataset (Akhtar et al., 2016), (b) ACTSA Sentiment Analysis corpus (Mukku and Mamidi, 2017), (c) IIT-Bombay Sentiment Analysis Dataset (A.R. et al., 2012), (d) BBC News Articles classification dataset, (e) iNLTK Headlines dataset, (f) Soham Bengali News classification corpus. The essential details of the datasets are described in Table 8.

Lang	Dataset	N	# Examples	
			Train	Test
hi	BBC Articles[7]	6	3,467	866
	IITP+ Movie Reviews	3	2,480	310
	IITP Product Reviews[8]	3	4,182	523
bn	Soham Articles[9]	6	11,284	1411
gu	iNLTK Headlines[10]	3	5,269	659
ml		3	5,036	630
mr		3	9,672	1,210
ta		3	5,346	669
te	ACTSA corpus[11]	3	4,328	541

Table 8: Statistics of publicly available datasets (N is the number of classes)

Some notes on the above mentioned public datasets

- The IITP+ Movie Reviews sentiment analysis dataset is created by merging IIT-Patna dataset with the smaller IIT-Bombay and iNLTK datasets.
- The IIT-Patna Movie and Product review datasets have 4 classes namely postive, negative, neutral and conflict. We ignored the conflict class.
- In the Telugu-ACTSA corpus, we evaluated only on the news line dataset (named as telugu_sentiment_fasttext.txt) and ignored all the other domain datasets as they have very few data-points.

[7] https://github.com/NirantK/hindi2vec/releases/tag/bbc-hindi-v0.1

[8] http://www.iitp.ac.in/ ai-nlp-ml/resources.html

[9] https://www.kaggle.com/csoham/classification-bengali-news-articles-indicnlp

[10] https://github.com/goru001/inltk

[11] https://github.com/NirantK/bharatNLP/releases