Findings of the First Shared Task on Indian Language Summarization (ILSUM): Approaches, Challenges and the Path Ahead

Shrey Satapara[1], Bhavan Modha[2], Sandip Modha[3] and Parth Mehta[4]

[1]Indian Institute of Technology, Hyderabad, India
[2]University Of Texas at Dallas, USA
[3]LDRP-ITR, Gandhinagar, India
[4]Parmonic, USA

Abstract
This paper provides an overview of the first edition of the shared task on Indian Language Summarization (ILSUM) organized at the 14th Forum for Information Retrieval Evaluation (FIRE 2022). The objective of this shared task was to create benchmark data for text summarization in Indian languages. This edition included three languages Hindi, Gujarati and Indian English. Indian English is an officially recognised dialect of English mainly used in the Indian subcontinent. The combined train and test datasets included more than 10000 article-summary pairs for each language which, to the best of our knowledge, is the largest publicly available summarization dataset for Indian languages. The task saw an enthusiastic response, with registrations from over 50 teams. A total of 13 teams submitted runs across the three languages out of which 10 teams submitted working notes. Standard ROUGE metrics were used as the evaluation metric. Indian English saw the most enthusiastic response with all 10 teams participating, followed by 6 teams submitting runs for Hindi with 5 teams for Gujarati.

Keywords
Automatic Text Summarization, Indian Languages, Headline Generation

1. Introduction

Research in Natural Language Processing has been known to be an uneven playing field for a long time. There is a chasm between the progress in resource-rich languages like English, Spanish, Chinese, etc as opposed to more resource-constrained languages like Hindi, Gujarati, Arabic, Urdu, etc. Although with the latest developments in the last few years, especially with open source large language models[1] and initiatives like the Forum for Information Retrieval Evaluation (FIRE)[2], this gap is slowly bridging. The progress however has been task-dependent. For instance tasks like hate speech detection[3, 4, 5, 6, 7], Sentiment analysis[8, 9], mixed script IR[10, 11], Indian legal document retrieval and summarization[12, 13, 14, 15, 16], Fake news detection[17, 18], authorship attribution[19, 20] to name a few, have made progress

Forum for Information Retrieval Evaluation, December 9-13, 2022, India
shreysatapara@gmail.com (S. Satapara); bhavanmodha@gmail.com (B. Modha); sjmodha@gmail.com (S. Modha); parth.mehta126@gmail.com (P. Mehta)
0000-0001-6222-1288 (S. Satapara); 0000-0002-1138-5968 (B. Modha); 0000-0003-2427-2433 (S. Modha); 0000-0002-4509-1298 (P. Mehta)

in past few years with several large scale datasets and pre-trained models becoming publicly available. Automatic text summarization on the other hand is one of the sub-disciplines of NLP where research is still more skewed towards English[21, 22, 23] and other resource-rich languages, while the focus on other resource-poor languages is almost negligible[24].

Indian languages, despite having millions of speakers, have received surprisingly little attention. While on one hand large-scale datasets with hundreds of thousands of documents exist for languages like English[25], Chinese[26], Spanish[27], etc., the datasets for any Indian language runs into at most a few dozen documents[28, 29, 30, 31, 32, 33]. Further most existing datasets are either not public or are too small to be useful. As a result, hardly any meaningful research has been possible in this area. Through this shared task, we aim to bridge the existing gap by creating reusable corpora for Indian Language Summarization.

In the first edition, we cover two major Indian languages Hindi and Gujarati, which have over 350 million and over 50 million speakers respectively. Apart from this we also include Indian English, a widely recognized dialect that can be substantially different from English spoken elsewhere. We provided over 10,000 news articles accompanied by a title and headlines for each language. Table 1 presents the details of the ILSUM dataset. The task is to generate a meaningful summary, either extractive or abstractive, for each article.

2. Related Work

The first serious attempt at creating a reusable dataset for automatic text summarization was perhaps made during the Document Understanding Conference (DUC)[34] in 2002. The dataset was a collection of news articles on 50 topics and four handwritten summaries for each article. This was followed up in subsequent years with new additions and new tasks. The DUC was later followed by the Text Analysis Conference (TAC)[35]. TAC ran for several years and, like DUC, produced several benchmark corpora. On the whole, the DUC and TAC datasets together have been by far the most popular datasets for evaluating text summarization. However, with the advent of deep learning and large language models, the DUC and TAC corpora became inadequate because of smaller corpus sizes. Since then the focus shifted to large-scale datasets that can be used for training deep neural networks. Often these datasets were built by collecting already available article summary pairs, for example from newspapers, rather than creating the summaries. One such very popular dataset is the CNN/Dailymail dataset[25]. The dataset consists of around 300K articles from CNN and Dailymail newspapers, and the headlines of the articles were used as a multi-sentence summary. This strategy was often reused for English as well as other languages. For instance, one of the largest Chinese datasets (LCTCS)[26] and Spanish (DACSA)[27] also employs the same strategy. A similar approach is also used for domain-specific summarizationParikh et al..

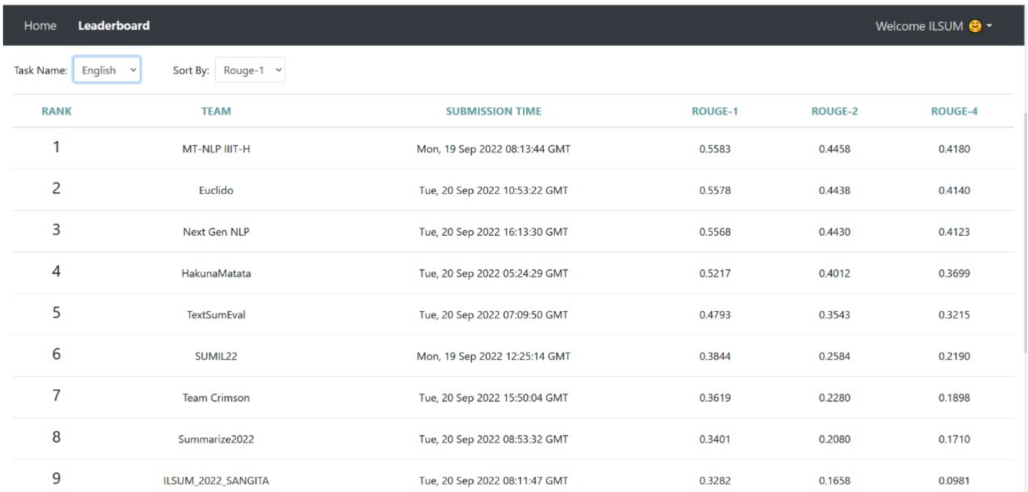
Compared to these the Indian language datasets are rather limited in size. Here we cover some of the more noteworthy attempts at creating text summarization datasets for Indian languages. An exhaustive list of the datasets is available made available in [24]. The most popular and cited corpus is a Malayalam dataset that was developed using news articles and human-written summary pairs[33]. The corpus has 100 documents and is not released publicly. It is mainly used by the same research group for experimentation and there are no reports from other groups that can validate the results. Another attempt is in the Bengali language that uses document summary pairs from printed NCTB books[29] but does not release the corpus publicly. The sole corpus for the Dogri language is also not public[32]. A corpus consisting of 71 folktales is the sole Konkani corpus[30] and has not been released publicly. A work on Sanskrit text summarization uses Wikipedia articles for the task[28]. However, the dataset is also not available publicly. A work on Kannada text summarization uses IR-based approaches but does not give details of the dataset used[31]. Overall, most if not all works on Indian language summarization do not have a public dataset and the works can not be substantiated by any studies that are independent of the original research papers.

3. Task Definition

The ILSUM task is a classic automatic summarization task where given a news article the participants are expected to generate a meaningful summary for the article. The summary can be either extractive or abstractive in nature. Traditionally the summarization tasks have been focused on generating a fixed-length summary irrespective of the input article length. This was especially the case with the DUC[34] and TAC[35] tasks and has since continued for a majority of the summarization tasks elsewhere. However, unlike DUC and TAC datasets where the length of the source articles and human generated summaries were controlled, this is not the case with more recent large scale corpora. If the source articles vary in length and informational content and so do the human summaries, forcing a fixed-length summary makes less sense.

Keeping this in mind we propose a different approach and do not attempt to generate a fixed-length summary. Instead, participants are expected to predict an appropriate summary length for each article and we only limit the maximum summary length to 75 words. We argue that too long or short length summary compared to the ground truth summary will adversely affect ROUGE precision or recall respectively and the F-measure will implicitly be penalized. For this task we use standard ROUGE metrics Rouge-1, Rouge-2 and Rouge-4 F-scores are used for evaluation.

To encourage participation and provide real time feedback a Kaggle like submission platform was provided to the participants. A separate leaderboard was provided for each language. During the validation phase, participants could submit runs on a blind validation dataset and instantly get the rouge scores. The leaderboard would display the highest score for each team along with the run id. During the test phase, participants could submit a maximum of three runs on the test data and see the rouge metrics instantly like in validation phase. The submission platform is shown below in figure 1



RANK	TEAM	SUBMISSION TIME	ROUGE-1	ROUGE-2	ROUGE-4
1	MT-NLP IIIT-H	Mon, 19 Sep 2022 08:13:44 GMT	0.5583	0.4458	0.4180
2	Euclido	Tue, 20 Sep 2022 10:53:22 GMT	0.5578	0.4438	0.4140
3	Next Gen NLP	Tue, 20 Sep 2022 16:13:30 GMT	0.5568	0.4430	0.4123
4	HakunaMatata	Tue, 20 Sep 2022 05:24:29 GMT	0.5217	0.4012	0.3699
5	TextSumEval	Tue, 20 Sep 2022 07:09:50 GMT	0.4793	0.3543	0.3215
6	SUMIL22	Mon, 19 Sep 2022 12:25:14 GMT	0.3844	0.2584	0.2190
7	Team Crimson	Tue, 20 Sep 2022 15:50:04 GMT	0.3619	0.2280	0.1898
8	Summarize2022	Tue, 20 Sep 2022 08:53:32 GMT	0.3401	0.2080	0.1710
9	ILSUM_2022_SANGITA	Tue, 20 Sep 2022 08:11:47 GMT	0.3282	0.1658	0.0981

Figure 1: ILSUM Submission Platform

4. Dataset

The dataset for this task is built using articles and headline pairs from several leading newspapers in the country. We have provided 10,000+ news articles for Hindi, 12000+ articles for Gujarati and 17900+ articles for Indian English. Table 1 shows the detail statistics of the train, test, and validation dataset. The task is to generate a meaningful fixed-length summary, either extractive or abstractive, for each article. While several previous works in other languages use news articles - headlines pair, the current dataset poses a unique challenge of code-mixing and script mixing. It is very common for news articles to borrow phrases from English, even if the article itself is written in an Indian Language. Examples like those shown below are a common occurrence both in the headlines as well as in the articles.

- Gujarati: "IND vs SA, 5મી T20 તસવીરોમાં: વરસાદે વિલન બની મજા બગાડી" (India vs SA, 5th T20 in pictures: rain spoils the match)

- Hindi: "LIC के IPO में पैसा लगाने वालों का टूटा दिल, आई एक और नुकसानदेह खबर" (Investors of LIC IPO left broken hearted, yet another bad news)

4.1. Dataset Creation

The news for ILSUM were scraped from the following news sites:

- www.indiatvnews.com(English)
- https://www.indiatv.in(Hindi)
- https://www.divyabhaskar.co.in(Gujarati)
- https://gujarati.news18.com(Gujarati)

The data was collected using a combination of web scraping tools beautifulsoup and Octoparse. We initially collected 19,839 English, 22,349 Gujarati, and 11,750 Hindi URLs. Next, we cleaned the data by removing the HTML codes and any additional junk like extra spaces. Further, we dropped the articles where the headlines were too short. Only articles where headline lenght was atleast 20 words were retained. The final corpus size is as shown in table 1

We assigned a unique id for each data record collected by computing a hash using the heading of the articles which are unique. The dataset was divided into train, test, and validation of size 70%(Train), 25%(Test) and 5%(Validation) respectively.

Table 1
Dataset Distribution

	Hindi	**Gujarati**	**English**
Training Set	6962	8460	12565
Validation Set	569	605	899
Test Set	2842	3021	4487
Total	10373	12086	17951

More details about the data are provided in table 2 below. The table contains number of sentences and words per article and per headline for all the three languages. It also shows number of codemixed articles (C.M.A.) and codemixed summaries(C.M.S.) for hindi and gujarati. As evident, english documents are the longest (in number of words), followed Hindi while Gujarati documents are the shortest. On the other hand, headlines are the longest in Hindi articles followed by English and Gujarati. There is a much higher level of codemixing in Gujarati articles compared to Hindi articles.

Table 2
Corpus Statistics

	Hindi			Gujarati			English		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
Sents/Article	17.13	17.78	17.73	23.41	23.78	22.99	19.23	19.54	19.58
Words/Article	407.41	422.68	421.45	369.35	375.74	364.73	487.33	494.76	495.98
Sents/Summary	1.6	1.61	1.63	1.29	1.17	1.18	1.27	1.3	1.28
Words/Summary	37	36.85	37.44	29.06	29.41	28.75	33.39	33.62	33.42
C.M.A.	248	30	125	286	32	91			
C.M.S.	363	25	100	2804	206	1011			

5. Methodology

In this section, we briefly discuss the approaches used by ILSUM 2022 participants. A majority of the teams preferred using large pre-trained models like BART, Pegasus, etc. for summarization and only a few approaches used traditional unsupervised methods. Notably, except for language-specific pre-trained models, none of the teams used any or language-specific resources, not even a stemmer or stopword list. This is counterintuitive in a task that would benefit widely from using linguistic resources. One possible reason is the lack of easy availability of such resources. Unlike for English, a limited number of resources exist for Hindi or Gujarati most of which are not well evaluated. This also gives us a pointer for the next version of ILSUM, which is to make these resources easily accessible and encouraging teams to use them. The summary of systems used by different teams for Hindi, Gujarati and English is described in table 3, 4 and 5

- **MT-NLP IIIT-H[36]:** Team MT-NLP-IITH achieved best performance in all three summarization tasks. The authors used various transformer models by fine-tuning and considering text summarization as a bottleneck task. For Hindi and Gujarati MT5, MBart, and IndicBART were finetuned for five epochs with a learning rate 5e-5 and max input length 512. Where best-performing model for Hindi is MT5 while MBart performed best for Gujarati. For English, PEGASUS, BART, T5 and ProphetNet were finetuned with similar hyperparameters, and PEGASUS outperformed other models on text data.
- **HakunaMatata[37]:** mT5 and IndicBART are fine-tuned with actual and augmented data of size five times bigger than actual data. Fine-tuned IndicBART outperformed mT5 on all three tasks.
- **Next Gen NLP[38]:** PEGASUS model worked best for this team on English and Gujarati where they use translation mapping-based approach. For hindi they used fine-tuned IndicBART model with augmented data.
- **PICT CL Lab[39]:** This team used a transformer-based abstract summary generation approach by Indic-BART based model, fine-tuned using language modelling loss.
- **TextSumEval[40]:** After preprocessing by removing multiple punctuations and emoticons, this team conducted four different experiments using LSTM, BART, GPT and T5 transformer, and T5 model achieved the best result for this team on English task.
- **SUMIL22[41]:** is one of the teams that use approaches other than pretrained LLMs. They calculate various text features such as sentence position, sentence length, sentence similarity, frequent words, and sentence numbers for each sentence. These text features and their optimized weights are used for sentence ranking, and then the summary is generated by selecting top-ranked sentences. The weight optimization of text features is done using the population-based meta-heuristic approach, Genetic Algorithm (GA).
- **Summarize2022[42]:** : For the English task, authors proposed a word frequency algorithm-based extractive text summarisation technique. Word frequency is calculated as the ratio of the frequency of a word and the frequency of the most occurring word in the text. Then sentence score is obtained by summing up the word frequency of all words occurring in a sentence. The mean of all sentence scores in the document is considered as a threshold to retain sentences in summary from the original text.
- **ILSUM_2022_SANGITA[43]:** The author proposed encoder-decoder-based architecture for the summarization task. Encoder Bi-LSTM has a hidden state dimension = 128; decoder lstm has a hidden dimension = 256. The word embedding size = 300. model is trained using rmsprop optimiser with sparse categorical cross-entropy loss for 50 epochs with a learning rate of Bart and batch size of 32.
- **IIIT_Ranchi[44]:** Extractive summarization approach using K means clustering was done by this team where clusters were created using sentence similarity scores. Where

no of clusters for a document containing fifteen sentences is six, five for a document containing six sentences and a document containing less than six sentences were left unmodified.

- **SSNCSENLP[45]:** mT5_m2m_CrossSum, a large-scale cross-lingual abstractive summarization model is used by this team to generate an abstractive summary.

Table 3
Methodology used for Hindi

Team Name	Method Description
MT-NLP IIIT-H[36]	MT5, MBart, and IndicBART. best in MT5
HakunaMatata[37]	MT5, and IndicBART with Data augmentation, best using IndicBART
Next Gen NLP[38]	Fine-tuned IndicBART Fine-tuned XL-Sum, best with IndicBART Fine-tuned mBART
PICT CL Lab 2[39]	Fine-tuned IndicBART
IIIT_Ranchi[44]	Extractive Summarization through K means clustring

Table 4
Methodology used for Gujarati

Team Name	Method Description
MT-NLP IIIT-H[36]	MT5[6], MBart[7] and IndicBART. best in MBart
HakunaMatata[37]	MT5, and IndicBART with Data augmentation
Next Gen NLP[38]	Translation Mapping with PEGASUS, Fine-tuned mBART, Fine-tuned XL-Sum. best Translation Mapping with PEGASUS
IIIT_Ranchi[44]	Extractive Summarization through K means clustring

Table 5
Methodology used on English Data

Team Name	Method Description
MT-NLP IIIT-H[36]	PEGASUS, BART, T5 and ProphetNet. PEGASUS gave best result
Next Gen NLP[38]	Fine-tuned PEGASUS Fine-tuned BRIO, SentenceBERT leveraged for summarization Fine-tuned T5
HakunaMatata[37]	MT5, and IndicBART with Data augmentation
TextSumEval[40]	LSTM based sequence-to-sequence model, BART model, GPT model, and T5 model, best with T5 Model
SUMIL22[41]	a population-based meta heuristic approach Genetic Algorithm
Summarize2022[42]	Word Frequency Algorithm
ILSUM_2022_SANGITA[43]	Bi-LSTM based encoder and LSTM Based Decoder
IIIT_Ranchi[44]	Extractive Summarization through K means clustering

6. Results

This section discusses results of runs submitted by different teams for all subtasks. Total of 12 teams submitted 47 runs across all subtasks. The summary of participation statistics is shown

Table 6
Participation Statistics

#Teams Registered	#Teams Submitted Runs	#Runs Submitted	#Paper Submitted
56	12	47	10

in Table 6. Table 7, 8 and 9 shows the performance of best runs submitted by each team on Hindi, Gujarati and English tasks, respectively.

Table 7
Performance of teams on Language summarization in Hindi

Rank	**Team Name**	**F1 Score**			
		ROUGE-1	**ROUGE-2**	**ROUGE-3**	**ROUGE-4**
1	MT-NLP IIIT-H[36]	0.607	0.510	0.484	0.471
2	HakunaMatata[37]	0.592	0.492	0.465	0.452
3	Euclido	0.583	0.480	0.452	0.439
4	Next Gen NLP[38]	0.556	0.455	0.427	0.414
5	PICT CL Lab 2[39]	0.544	0.443	0.419	0.400
6	IIIT_Ranchi[44]	0.327	0.174	0.136	0.126
Late Entry	SSNCSENLP[45]	0.379	0.225	0.170	0.135

Table 8
Performance of teams on Language summarization in Gujarati

Rank	**Team Name**	**F1 Score**			
		ROUGE-1	**ROUGE-2**	**ROUGE-3**	**ROUGE-4**
1	MT-NLP IIIT-H[36]	0.261	0.165	0.138	0.124
2	HakunaMatata[37]	0.243	0.146	0.119	0.106
3	Euclido	0.225	0.123	0.091	0.075
4	Next Gen NLP[38]	0.209	0.119	0.095	0.084
5	IIIT_Ranchi[44]	0.176	0.085	0.063	0.053

Some of the summaries generated by the participating teams are listed alongside the gold-standard summaries below. Some of the summaries are codemixed and use one or two english words besides using english numerals. The quality of code-mixed summaries generated by the participating teams are at par with single script summaries.

Hindi

- ***Original***: हिमाचल प्रदेश: Flash Flood की वजह से नाले में अचानक बढ़ा पानी, 1 की मौत, 9 लापता",लाहौल स्पिति के एसपी मानव वर्मा ने बताया कि लाहौल स्पिति की उदयपुर डिवीजन में फ्लैश फ्लड की वजह से 9 लोग लापता हैं।
- ***MT-NLP IIIT-H***: IANS द्वारा दी गई सूचना के अनुसार, आपदा मनाली–लेह राजमार्ग पर स्थित उदयपुर में हुई और तोजिंग नदी पर एक महत्वपूर्ण पुल क्षतिग्रस्त हो गया। रिपोर्ट्स में कहा गया है कि पर्यटकों सहित कई वाहन राजमार्ग पर फंस गए हैं।

Table 9
Performance of teams on Language summarization in English

Rank	Team Name	F1 Score			
		ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
1	MT-NLP IIIT-H[36]	0.558	0.446	0.426	0.418
2	Euclido	0.558	0.444	0.422	0.414
3	Next Gen NLP[38]	0.557	0.443	0.421	0.412
4	HakunaMatata[37]	0.522	0.401	0.379	0.370
5	TextSumEval[40]	0.479	0.354	0.330	0.322
6	SUMIL22[41]	0.384	0.258	0.231	0.219
7	Team Crimson	0.362	0.228	0.201	0.190
8	Summarize2022[42]	0.340	0.208	0.181	0.171
9	ILSUM_2022_SANGITA[43]	0.328	0.166	0.122	0.098
10	IIIT_Ranchi[44]	0.321	0.186	0.163	0.154

- ***HakunaMatata***: लाहौल स्पिति के एसपी मानव वर्मा ने बताया कि लाहौल स्पिति की उदयपुर डिवीजन में फ्लैश फ्लड की वजह से 9 लोग लापता हैं।
- ***Euclido***: हिमाचल प्रदेश के लाहौल स्पिति में अचानक आई Flash Flood की वजह से तोजिंग नाले (Tozing Nullah) में पानी बढ़ गया, जिस वजह से एक व्यक्ति की मौत हो गई, एक घायल है और 9 लोग लापता है।

Gujarati

- ***Original***: Ahmedabad Airport: ધુળેટીના તહેવારને લઈ અમદાવાદ એરપોર્ટ મેઘધનુષી રંગોમાં રંગાયું,"અમદાવાદના સરદાર વલ્લભભાઈ પટેલ ઈન્ટરનેશનલ એરપોર્ટ (Sardar Vallabhbhai Patel International Airport, Ahmedabad) પર રંગોત્સવ હોળી (Holi 20222) ના તહેવારને મનાવવાનો થનગનાટ ઉડીને આંખે વળગી રહ્યો છે. મુસાફરોને હોળી (Holi Festival) ના મેઘધનુષી રંગોમાં તરબોળ કરવા એરપોર્ટ (Ahmedabad Airport) પરિસરમાં વિવિધ પ્રવૃત્તિઓનું આયોજન કરવામાં આવ્યું છે.
- ***MT-NLP IIIT-H***: મુસાફરોને વરલી અને માંડલ જેવી પરંપરાગત કળાના વિવધ સ્વરૂપોથી માહિતગાર કરવાનો પ્રયાસ કરવામાં આવ્યો છે. દેશ–વિદેશના પ્રવાસીઓ અહીં વિવિધ કળાને માણવાનો અનુભવ સેલ્ફીમાં પણ કેદ કરી રહ્યા છે.
- ***HakunaMatata***: SVPI એરપોર્ટ પર ભારતીય કળાની વૈવિધ્યસભરતાનું આબેહૂબ નિદર્શન કરાવવામાં આવ્યું છે. મુસાફરો રંગબેરંગી માહોલમાં મંડલા અને વરલી જેવી પરંપરાગત કળાનો અદભૂત અનુભવ પણ માણી શકે તેવુ આયોજન કરવામાં આવ્યું છે.
- ***Euclido***: મુસાફરો રંગબેરંગી માહોલમાં મંડલા અને માંડલ જેવી પરંપરાગત કળાના વિવધ સ્વરૂપોથી માહિતગાર કરવાનો પ્રયાસ કરવામાં આવ્યો છે

7. Conclusion and Future Work

The Indian Language Summarization (ILSUM) track at FIRE'22 is the first attempt to create benchmarked corpora for text summarization of Indian languages such as Hindi and Gujarati in addition to English. The majority of the summarization systems, submitted by the various participants, were based on pre-trained models like MT5, MBart, and IndicBART. Some of the participants also submitted systems using traditional unsupervised approaches, such as TexRank. The reported evaluation metric, the rouge F-Score, was comparable between English and Hindi corpora but significantly lower in Gujarati corpora. In the next edition of the ILSUM, we are planning to create a similar corpus for other languages like Bengali and Dravidian languages like Tamil and Telugu.

References

[1] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://doi.org/10.18653/v1/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[2] P. Mehta, T. Mandl, P. Majumder, S. Gangopadhyay, Report on the FIRE 2020 evaluation initiative, SIGIR Forum 55 (2021) 3:1–3:11. URL: https://doi.org/10.1145/3476415.3476418. doi:10.1145/3476415.3476418.

[3] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hatespeech and offensive content identification in english and indo-aryan languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1–19. URL: http://ceur-ws.org/Vol-3159/T1-1.pdf.

[4] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 87–111. URL: http://ceur-ws.org/Vol-2826/T2-1.pdf.

[5] S. Modha, T. Mandl, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 167–190. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[6] H. Madhu, S. Satapara, S. Modha, T. Mandl, P. Majumder, Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments, Expert Systems with Applications (2022) 119342.

[7] S. Modha, P. Majumder, T. Mandl, C. Mandalia, Detecting and visualizing hate speech in

social media: A cyber watchdog for surveillance, Expert Syst. Appl. 161 (2020) 113725. URL: https://doi.org/10.1016/j.eswa.2020.113725. doi:10.1016/j.eswa.2020.113725.

[8] M. Subramanian, R. Ponnusamy, S. Benhur, K. Shanmugavadivel, A. Ganesan, D. Ravi, G. K. Shanmugasundaram, R. Priyadharshini, B. R. Chakravarthi, Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer, Comput. Speech Lang. 76 (2022) 101404. URL: https://doi.org/10.1016/j.csl.2022.101404. doi:10.1016/j.csl.2022.101404.

[9] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, B. Premjith, S. K, S. C. Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the hasoc-dravidiancodemix shared task on offensive language detection in tamil and malayalam, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 589–602. URL: http://ceur-ws.org/Vol-3159/T3-1.pdf.

[10] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, M. Choudhury, Overview of the mixed script information retrieval (MSIR) at FIRE-2016, in: P. Majumder, M. Mitra, P. Mehta, J. Sankhavara, K. Ghosh (Eds.), Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, volume 1737 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 94–99. URL: http://ceur-ws.org/Vol-1737/T3-1.pdf.

[11] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, P. Rosso, Query expansion for mixed-script information retrieval, in: S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, K. Järvelin (Eds.), The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, ACM, 2014, pp. 677–686. URL: https://doi.org/10.1145/2600428.2609622. doi:10.1145/2600428.2609622.

[12] P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya, P. Majumder, Overview of the FIRE 2019 AILA track: Artificial intelligence for legal assistance, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 1–12. URL: http://ceur-ws.org/Vol-2517/T1-1.pdf.

[13] P. Bhattacharya, P. Mehta, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, FIRE 2020 AILA track: Artificial intelligence for legal assistance, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp. 1–3. URL: https://doi.org/10.1145/3441501.3441510. doi:10.1145/3441501.3441510.

[14] V. Parikh, U. Bhattacharya, P. Mehta, A. Bandyopadhyay, P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, AILA 2021: Shared task on artificial intelligence for legal assistance, in: D. Ganguly, S. Gangopadhyay, M. Mitra, P. Majumder (Eds.), FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021, ACM, 2021, pp. 12–15. URL: https://doi.org/10.1145/3503162.3506571. doi:10.1145/3503162.3506571.

[15] V. Parikh, V. Mathur, P. Mehta, N. Mittal, P. Majumder, Lawsum: A weakly supervised

approach for indian legal document summarization, CoRR abs/2110.01188 (2021). URL: https://arxiv.org/abs/2110.01188. arXiv:2110.01188.
[16] S. Ghosh, A. Wyner, Identification of rhetorical roles of sentences in indian legal judgments, in: Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference, volume 322, IOS Press, 2019, p. 3.
[17] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake news detection in the urdu language, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 2537–2542. URL: https://aclanthology.org/2020.lrec-1.309/.
[18] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. F. Gelbukh, Threatening language detection and target identification in urdu tweets, IEEE Access 9 (2021) 128302–128313. URL: https://doi.org/10.1109/ACCESS.2021.3112500. doi:10.1109/ACCESS.2021.3112500.
[19] P. Mehta, P. Majumder, Optimum parameter selection for K.L.D. based authorship attribution in gujarati, in: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013, Asian Federation of Natural Language Processing / ACL, 2013, pp. 1102–1106. URL: https://aclanthology.org/I13-1155/.
[20] P. Mehta, P. Majumder, Large scale quantitative analysis of three indo-aryan languages, J. Quant. Linguistics 23 (2016) 109–132. URL: https://doi.org/10.1080/09296174.2015.1071151. doi:10.1080/09296174.2015.1071151.
[21] P. Mehta, From extractive to abstractive summarization: A journey, in: H. He, T. Lei, W. Roberts (Eds.), Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany, August 7-12, 2016, Association for Computational Linguistics, 2016, pp. 100–106. URL: https://doi.org/10.18653/v1/P16-3015. doi:10.18653/v1/P16-3015.
[22] P. Mehta, P. Majumder, Effective aggregation of various summarization techniques, Inf. Process. Manag. 54 (2018) 145–158. URL: https://doi.org/10.1016/j.ipm.2017.11.002. doi:10.1016/j.ipm.2017.11.002.
[23] S. Modha, P. Majumder, T. Mandl, R. Singla, Design and analysis of microblog-based summarization system, Social Network Analysis and Mining 11 (2021) 1–16. URL: https://doi.org/10.1007/s13278-021-00830-3.
[24] S. Sinha, G. N. Jha, An overview of indian language datasets used for text summarization, CoRR abs/2203.16127 (2022). URL: https://doi.org/10.48550/arXiv.2203.16127. doi:10.48550/arXiv.2203.16127. arXiv:2203.16127.
[25] R. Nallapati, B. Zhou, C. N. dos Santos, Ç. Gülçehre, B. Xiang, Abstractive text summarization using sequence-to-sequence rnns and beyond, in: Y. Goldberg, S. Riezler (Eds.), Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016, ACL, 2016, pp. 280–290. URL: https://doi.org/10.18653/v1/k16-1028. doi:10.18653/v1/k16-1028.
[26] B. Hu, Q. Chen, F. Zhu, LCSTS: A large scale chinese short text summarization dataset, in: L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics, 2015, pp.

1967–1972. URL: https://doi.org/10.18653/v1/d15-1229. doi:10.18653/v1/d15-1229.
[27] E. S. Soriano, V. Ahuir, L. Hurtado, J. González, DACSA: A large-scale dataset for automatic summarization of catalan and spanish newspaper articles, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 5931–5943. URL: https://doi.org/10.18653/v1/2022.naacl-main.434. doi:10.18653/v1/2022.naacl-main.434.
[28] S. Barve, S. Desai, R. Sardinha, Query-based extractive text summarization for sanskrit, in: S. Das, T. Pal, S. Kar, S. C. Satapathy, J. K. Mandal (Eds.), Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications, FICTA 2015, Durgapur, India, 16-18 November 2015, volume 404 of *Advances in Intelligent Systems and Computing*, Springer, 2015, pp. 559–568. URL: https://doi.org/10.1007/978-81-322-2695-6_47. doi:10.1007/978-81-322-2695-6_47.
[29] R. R. Chowdhury, M. T. Nayeem, T. T. Mim, M. S. R. Chowdhury, T. Jannat, Unsupervised abstractive summarization of bengali text documents, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, Association for Computational Linguistics, 2021, pp. 2612–2619. URL: https://doi.org/10.18653/v1/2021.eacl-main.224. doi:10.18653/v1/2021.eacl-main.224.
[30] J. D'Silva, U. Sharma, Development of a konkani language dataset for automatic text summarization and its challenges, International Journal of Engineering Research and Technology. International Research Publication House. ISSN (2019) 0974–3154.
[31] V. R. Embar, S. R. Deshpande, A. Vaishnavi, V. Jain, J. S. Kallimani, saramsha-a kannada abstractive summarizer, in: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2013, pp. 540–544.
[32] S. Gandotra, B. Arora, Feature selection and extraction for dogri text summarization, in: Rising Threats in Expert Applications and Solutions, Springer, 2021, pp. 549–556.
[33] R. Kabeer, S. M. Idicula, Text summarization for malayalam documents - an experience, in: International Conference on Data Science & Engineering, ICDSE 2014, Kochi, India, August 26-28, 2014, IEEE, 2014, pp. 145–150. URL: https://doi.org/10.1109/ICDSE.2014.6974627. doi:10.1109/ICDSE.2014.6974627.
[34] A. Nenkova, Automatic text summarization of newswire: Lessons learned from the document understanding conference, in: M. M. Veloso, S. Kambhampati (Eds.), Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA, AAAI Press / The MIT Press, 2005, pp. 1436–1441. URL: http://www.aaai.org/Library/AAAI/2005/aaai05-228.php.
[35] H. T. Dang, K. Owczarzak, Overview of the TAC 2008 update summarization task, in: Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008, NIST, 2008. URL: https://tac.nist.gov/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf.
[36] A. Urlana, S. M. Bhatt, N. Surange, M. Shrivastava, Indian Language Summarization using Pretrained Sequence-to-Sequence Models, in: Working Notes of FIRE 2022 - Forum for

Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[37] D. Taunk, V. Varma, Summarizing Indian Languages using Multilingual Transformers based Models, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[38] R. Tangsali, A. Pingle, A. Vyawahare, I. Joshi, R. Joshi, Implementing Deep Learning-Based Approaches for Article Summarization in Indian Languages, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[39] A. Agarwal, S. Naik, S. Sonawane, Abstractive Text Summarization for Hindi Language using IndicBART, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[40] S. Chakraborty, D. Kaushik, S. R. Laskar, P. Pakray, Exploring Text Summarization Models for Indian Languages Summarization, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[41] V. P. K. Doppalapudi, S. R. Srigadha, P. Verma, S. Pal, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[42] N. Abinaya, S. Anbukkarasi, S. Varadhaganapathy, An Extractive Text Summarization Using Word Frequency Algorithm for English Text, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[43] S. Singh, J. P. Singh, A. Deepak, Deep Learning based Abstractive Summarization for English Language, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
[44] K. Kumari, R. Kumari, An Extractive Approach for Automated Summarization of Indian Languages using Clustering Techniques, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[45] A. Krishnakumar, F. Naushin, M. KL, B. B, Text summarization for Indian languages using pre-trained models, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022.