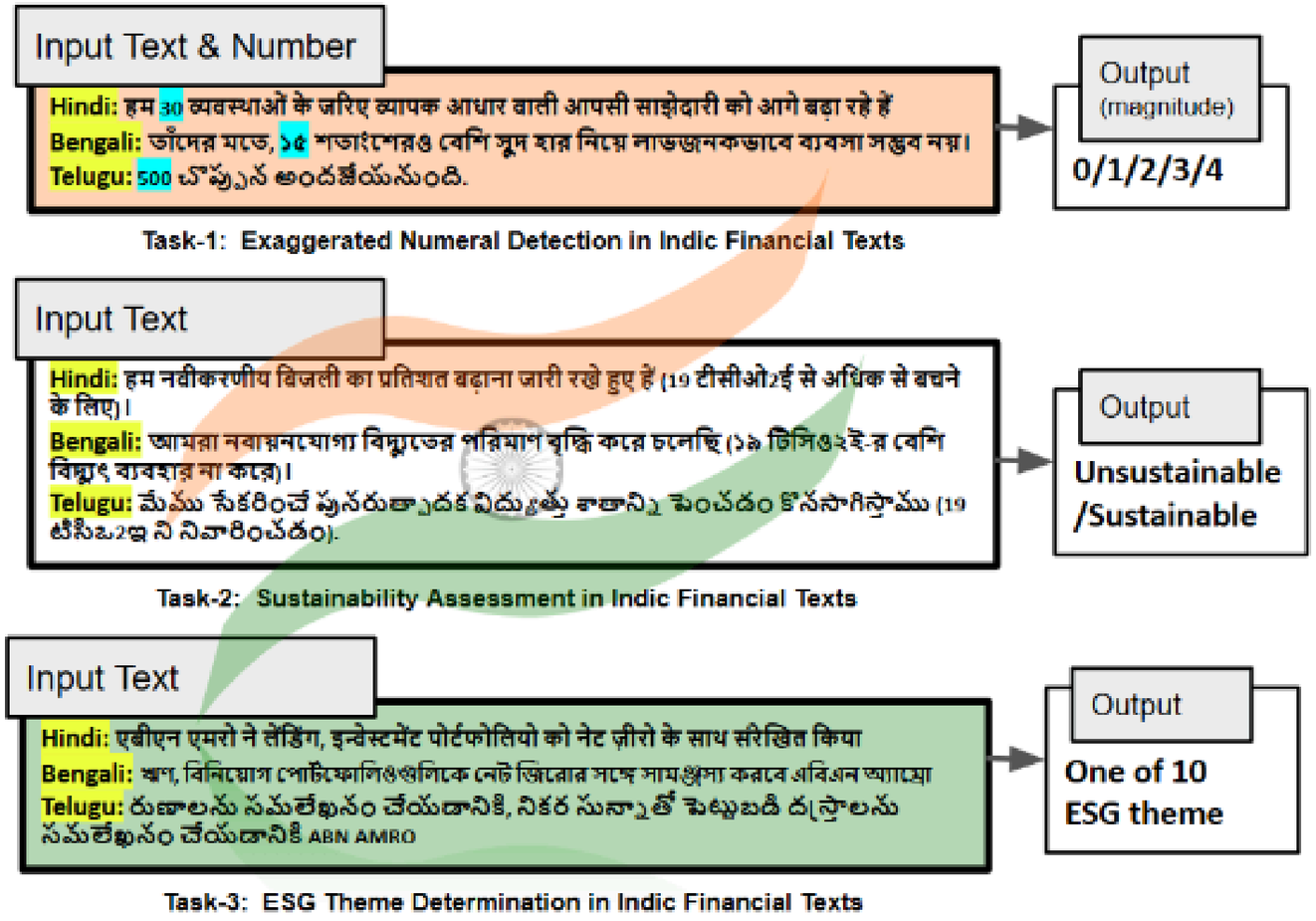
INDICFINNLP: FINANCIAL NATURAL LANGUAGE PROCESSING FOR INDIAN LANGUAGES

LREC-COLING 2024

TORINO, ITALIA

Sohom Ghosh, Arnab Majhi, Aswartha Narayana, Sudip Kumar Naskar

PROBLEM STATEMENT



Languages: Hindi, Bengali, Telugu

Tasks: Exaggerated Numeral Detection, Sustainability Assessment, ESG Theme Determination

EXAGGERATED NUMERAL DETECTION DATASET

Source:

- Budget speeches of Hindi, Bengali, and Telugu-speaking states (Punjab, Uttarakhand, Haryana, West Bengal, Telangana, and Andhra Pradesh) starting from the year 2011 till 2023
- Financial texts filtered from the Samanantar corpus (Rameshet al., TACL 2022)

Language	0	1	2	3	4
Hindi	2435	3624	1444	2485	652
Bengali	1574	1886	931	1416	323
Telugu	1737	1800	983	1182	314

Table 1: Task-1 label wise distribution. 0/1/2/3/4 are the magnitudes

indic	number_english	number_indic	start_posn	end_posn	language	magnitude
মাতিল ম্যানুকান (১৯১৪-২০০১), রিয়েল এস্টেট বিনিয়োগকারী।	2001	২০০১	22	26	bengali	3

SUSTAINABILITY ASSESSMENT DATASET

Source:

- Translated the existing dataset proposed by Kangand El Maarouf (FinSim4-ESG FinNLP-2022) from English to Indian languages (Hindi, Bengali, and Telugu). Retained only the high quality ones.

Language	BS(F1)	Sim.	Class	#
Hindi	>= 0.90	>=0.75	S	1212
			US	1026
Bengali	>=0.88	>=0.68	S	1203
			US	1025
Telugu	>=0.88	>=0.80	S	1119
			US	953

Table 2: Task-2 data distribution & thresholds. S=Sustainable, U=Unsustainable. BS=BERTScore, Sim.=Cosine Similarity

sentence_indic	label	language
आपके संगठन का सकल वैश्विक स्कोप 1 उत्सर्जन मीट्रिक टन में कितना था?	unsustainable	hindi

ESG THEME DETERMINATION DATASET

Source:

- Translated the existing dataset proposed by Chen et al. 2023 (FinNLP-2023 ML-ESG) from English to Indian languages (Hindi, Bengali, and Telugu). Manually verified and corrected the translations wherever needed.

ESG Theme	#
climate change	92
corporate governance	91
environmental opportunities	72
product liability	68
natural capital	50
pollution waste	44
human capital	37
corporate behavior	30
social opportunities	27
stake holder opposition	21

Table 3: Task-3 label wise distribution for Hindi, Bengali, and Telugu.

URL	news_title_indic	ESG_theme	language
https://www.esgtoday.com/abn-amro-to-align-lending-investment-portfolios-with-net-zero/	రుణాలను సమలేఖనం చేయడానికి, నికర సున్నాతో పెట్టుబడి దస్త్రాలను సమలేఖనం చేయడానికి ABN AMRO	climate chnage	telugu

RESULTS: EXAGGERATED NUMERAL DETECTION

Ts	L	Model	Test			
			Pr	Re	F1	Acc
1	H	MB+LGB	0.63	0.64	0.63	0.64
1	H	IB+LGB	0.44	0.49	0.45	0.49
1	H	MB+XGB	0.63	0.64	0.63	0.64
1	H	IB+XGB	0.46	0.49	0.46	0.49
1	H	MB+SVM	**0.69**	**0.68**	**0.68**	**0.68**
1	B	MB+LGB	0.64	0.64	0.63	0.64
1	B	IB+LGB	0.51	0.51	0.50	0.51
1	B	MB+XGB	0.62	0.62	0.61	0.62
1	B	IB+XGB	0.51	0.50	0.48	0.50
1	B	MB+SVM	**0.66**	**0.65**	**0.65**	**0.65**
1	T	MB+LGB	0.59	0.61	0.59	0.61
1	T	IB+LGB	0.44	0.46	0.44	0.46
1	T	MB+XGB	0.59	0.60	0.59	0.60
1	T	IB+XGB	0.41	0.43	0.41	0.43
1	T	IB+XGB	**0.69**	**0.68**	**0.68**	**0.68**

Table 4: Tasks (Ts) 1, 2, 3 results for Languages (L) Hindi (H), Bengali (B), Telugu (T). E=English, -P= -Paraphrased, IB=IndicBERT, MB=MBERT, XGB=XGBoost, LGB=LightGBM, Pr=Precision, Re=Recall, Acc=Accuracy. **Bold** means the best.

RESULTS: SUSTAINABILITY ASSESSMENT

Ts	L	Model	Test			
			Pr	Re	F1	Acc
2	H	IB	0.86	0.86	0.86	0.86
2	H	MB	0.77	0.77	0.77	0.77
2	H	MLM-IB	0.29	0.54	0.38	0.54
2	H	E-Ro	**0.95**	**0.95**	**0.95**	**0.95**
2	B	IB	0.80	0.80	0.80	0.80
2	B	MB	0.76	0.76	0.76	0.76
2	B	MLM-IB	0.81	0.81	0.81	0.81
2	B	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**
2	T	IB	0.79	0.79	0.79	0.78
2	T	MB	0.90	0.89	0.89	0.89
2	T	MLM-IB	0.90	0.90	0.90	0.90
2	T	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**

Table 4: Tasks (Ts) 1, 2, 3 results for Languages (L) Hindi (H), Bengali (B), Telugu (T). E=English, -P= -Paraphrased, IB=IndicBERT, MB=MBERT, XGB=XGBoost, LGB=LightGBM, Pr=Precision, Re=Recall, Acc=Accuracy. **Bold** means the best.

RESULTS: ESG THEME DETERMINATION

Ts	L	Model	Test			
			Pr	Re	F1	Acc
3	H	IB	0.03	0.17	0.05	0.17
3	H	MB	0.20	0.20	0.08	0.20
3	H	MLM-MB	0.20	0.20	0.11	0.20
3	H	E-MB	0.11	0.30	0.16	0.30
3	H	IB-P	0.12	0.26	0.16	0.26
3	H	MB-P	0.45	0.48	0.44	0.48
3	H	MLM-MB-P	0.43	0.46	0.44	0.46
3	H	E-MB-P	**0.56**	**0.63**	**0.59**	**0.63**
3	B	IB	0.03	0.17	0.05	0.17
3	B	MB	0.03	0.17	0.05	0.17
3	B	MLM-MB	0.11	0.20	0.10	0.20
3	B	E-MB	0.11	0.26	0.14	0.26
3	B	IB-P	0.20	0.30	0.23	0.30
3	B	MB-P	0.40	0.37	0.35	0.37
3	B	MLM-IB-P	0.32	0.37	0.33	0.37
3	B	E-MB-P	**0.55**	**0.59**	**0.55**	**0.59**
3	T	IB	0.03	0.17	0.05	0.17
3	T	MB	0.09	0.24	0.12	0.24
3	T	MLM-MB	0.07	0.22	0.11	0.22
3	T	E-MB	0.07	0.22	0.11	0.22
3	T	IB-P	0.27	0.31	0.22	0.31
3	T	MB-P	0.44	0.46	0.42	0.46
3	T	MLM-MB-P	0.36	0.41	0.37	0.41
3	T	E-MB-P	**0.56**	**0.63**	**0.58**	**0.63**

Table 4: Tasks (Ts) 1, 2, 3 results for Languages (L) Hindi (H), Bengali (B), Telugu (T). E=English, -P= -Paraphrased, IB=IndicBERT, MB=MBERT, XGB=XGBoost, LGB=LightGBM, Pr=Precision, Re=Recall, Acc=Accuracy. **Bold** means the best.