INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIIT Hyderabad

LREC-COLING 2024

TeClass: A Human Annotated Relevance-based Headline Classification and Generation Dataset for Telugu

Gopichand Kanumolu*, Lokesh Madasu*, Nirmal Surange, Manish Shrivastava

Language Technologies Research Center, IIIT-Hyderabad

Introduction

Headline Generation

- **Task:** Generating a relevant headline that represents the core information present in the news article

- **Challenges:**
  - The performance of the headline generation model depends on the quality of the training data
  - The presence of irrelevant headlines in news articles scraped from the web often results in sub-optimal performance

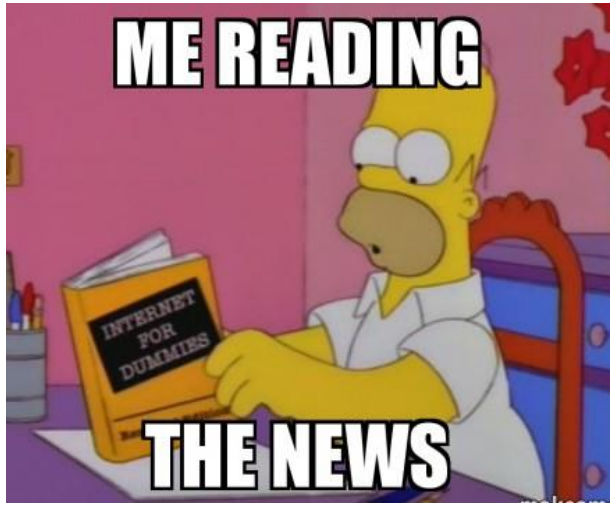
- **Solution:**
  - We propose that relevance-based headline classification can greatly aid the task of generating relevant headlines.

Introduction

Relevance-based Headline Classification

- Categorizing a news headline based on its relevance to the corresponding news article
  - **Highly Relevant**
  - **Moderately Relevant**
  - **Least Relevant**

- Important task in assessing the relationship between headline and its article, can be useful for applications including:
  - News Recommendation
  - Incongruent Headline Detection
  - Headline Stance Classification



My reaction, if the relevance is

High Moderate Least

Key Contributions

Headline Classification

• We present "TeClass", a large, diverse, and high-quality human annotated dataset for Telugu

• It contains 26,178 article-headline pairs annotated for relevance-based headline classification with one of the three categories:

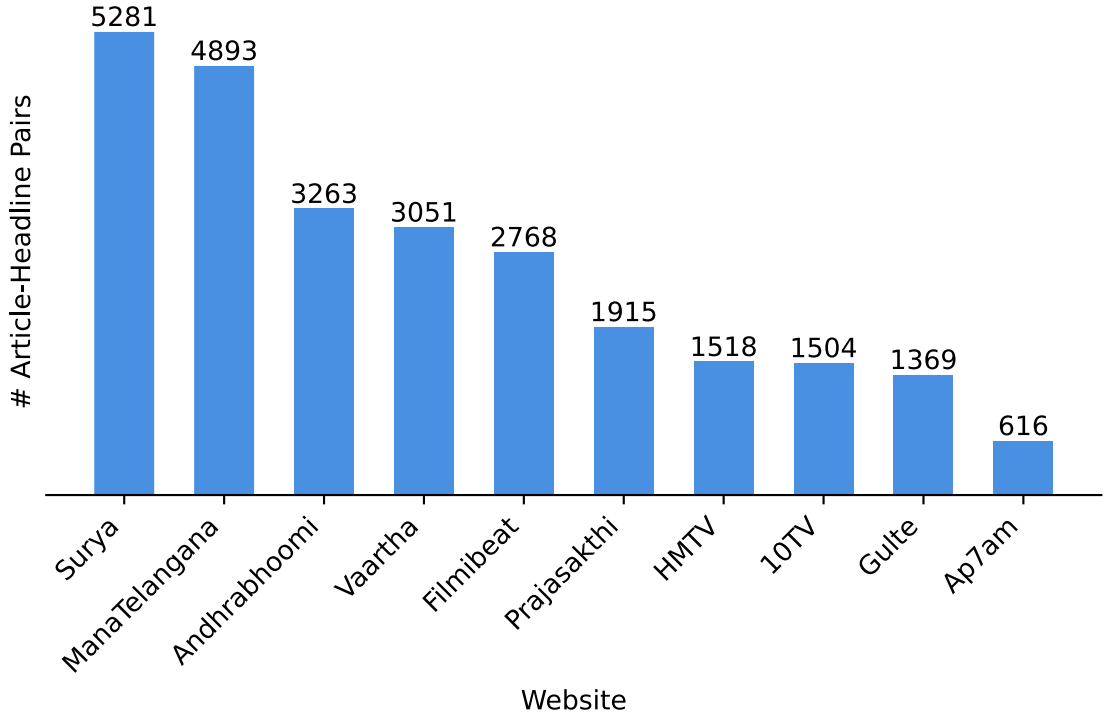
- ❑ Highly Related (HREL)
- ❑ Moderately Related (MREL)
- ❑ Least Related (LREL)

Headline Generation

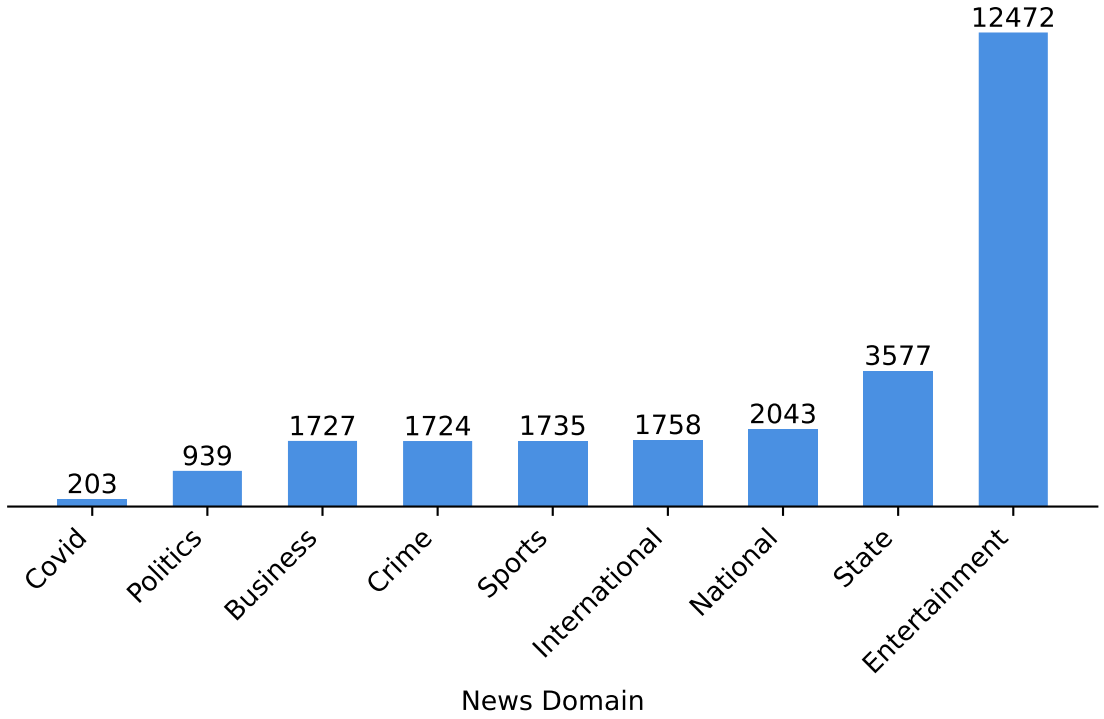
- We study the impact of fine-tuning headline generation models on different types of headlines (with varying degrees of relevance to the article)
- We demonstrate that the task of relevant headline generation is best served when the generation models are fine-tuned on highly relevant data even if the highly relevant article-headline pairs are significantly less in number.

TeClass Dataset Creation: Selecting Data for Annotation

- Developed site-specific web scrapers to collect the article-headline pairs from multiple news websites



- Covered a broad spectrum of domains, including State, National, International, Sports, Business, Politics, Entertainment, Crime, and Covid



TeClass Dataset Creation: Annotation Guidelines

Highly Related

Factual Main Event (FME): The headline is mostly explicitly present in the article and represents the main event addressed in the article which is factually correct.

Article: మంత్రి తానేటి వనిత సంతకం ఫోర్జరీ చేశారు . మంత్రి సంతకాన్ని కడప జిల్లాకు చెందిన టీడీపీ నేత ఫోర్జరీ చేశాడు. మంత్రి తానేటి వనిత సంతకం లెటర్ ప్యాడ్ పై ఫోర్జరీ చేశారు . అసైన్డ్ భూమి కేటాయించాలని కలెక్టర్ కి టీడీపీ నేత నకిలీ లేఖ ఇచ్చాడు . మంత్రి సంతకం ఫోర్జరీ చేసి టీడీపీ నేత దొరికిపోయాడు . మంత్రి తానేటి వనిత తన సంతకం ఫోర్జరీపై డిజిపికి పిర్యాదు చేసింది . సంతకం ఫోర్జరీ చేసిన వారిపై కఠిన చర్యలు తీసుకోవాలని పిర్యాదు చేసింది .

Translation: Minister Taneti Vanitha's signature was forged. The minister's signature was forged by a TDP leader from Kadapa district. Minister Thaneti Vanitha's signature was forged on the letterpad. The TDP leader had given a fake letter to the collector asking him to allot the assigned land. The TDP leader was caught for forging the signature of the minister. Minister Thaneti Vanitha had lodged a complaint with the DGP over the forgery of her signature. She has also filed a complaint seeking strict action against those who forged the signature.

Headline: మంత్రి తానేటి వనిత సంతకం ఫోర్జరీ

Translation: Minister Taneti Vanitha's signature forged

Explanation: The main event being discussed in the article is the forgery of the signature of minister Taneti Vanitha. The headline also presents the same information.

TeClass Dataset Creation: Annotation Guidelines

Moderately Related

i. **Strong Conclusion (STC):** The headline is not explicitly present (in the same words) in the article, but it can be inferred from the article .

ii. **Factual Secondary Event (FSE):** The headline represents a secondary event addressed in the article which is factually correct.

iii. **Weak Conclusion (WKC):** The headline is not explicitly present (in the same words) in the article, and it has been inferred from only a small portion of the article.

Article: అమరావతి : రెండు తెలుగు రాష్ట్రాల మధ్య జల వివాదం ఏర్పడిన నేపథ్యంలో కృష్ణా , గోదావరి నదీ జలాల బోర్డుల పరిధులను ఖరారుచేస్తూ మొన్న అర్ధరాత్రి కేంద్ర జలశక్తి మంత్రిత్వ శాఖ గెజిట్ను విడుదల చేసిన విషయం తెలిసిందే. దీనిపై టీడీపీ అధినేత చంద్రబాబు నాయుడు స్పందించారు. ఆ గెజిట్పై పూర్తిగా అధ్యయనం చేశాకే స్పందిస్తానని అన్నారు. విజయవాడలోని రమేశ్ ఆసుపత్రికి వెళ్లి అక్కడ చికిత్స పొందుతున్న ఎమ్మెల్సీ బచ్చుల అర్జునుడుని చంద్రబాబు పరామర్శించి అనంతరం మీడియాతో మాట్లాడుతూ .. బచావత్ ట్రైబ్యునల్కు , గెజిట్కు ఉన్న వ్యత్యాసాలను గుర్తించాల్సి ఉందని ఆయన అన్నారు. అయితే , ఈ విషయాలను ప్రస్తావించకుండా వైస్సార్సీపీ ప్రభుత్వం తప్పించుకునే ప్రయత్నం చేస్తోందని వివమర్శించారు. ఏపీ పట్ల సీఎం జగన్ బాధ్యత లేకుండా వ్యవహరిస్తున్నారని , తాము మాత్రం ఏపీ ప్రయోజనాల కోసం పోరాడతూనే ఉంటామని ఆయన చెప్పుకొచ్చారు.

Translation: Amaravati: In the wake of the water dispute between the two Telugu states, the Union Jal Shakti Ministry has released a gazette notification finalising the limits of the Krishna and Godavari river water boards. On this, the TDP chief Chandrababu Naidu responded. He said he would respond only after a thorough study of the gazette. Chandrababu went to the Ramesh Hospital in Vijayawada and visited MLC Bachula Arjunudu, who is undergoing treatment there, and later spoke to the media. He said the differences between the Bachawat Tribunal and the Gazette need to be identified. However, he said that the YSRCP government was trying to avoid mentioning these issues. He said that CM Jagan is acting irresponsibly towards AP and they will continue to fight for the interests of AP.

Headline: ఏపీ ప్రయోజనాల కోసం పోరాడతూనే ఉంటాం

Translation: We will continue to fight for the interests of AP

Explanation: The article mainly focuses on Chandrababu Naidu's reaction to the Gazette published by the Central Ministry of Jal Shakti. However, the headline only reflects a small portion of the article that discusses his statement, "We will fight for the benefits of AP."

TeClass Dataset Creation: Annotation Guidelines

Least Related

i. **Sensational (SEN):** The headline is intended to catch the attention of the reader, by reporting emotionally loaded impressions/controversial statements that manipulate the truth of the story.

ii. **Clickbait (CBT):** A headline that tempts the reader to click on the link, where there is an extreme disconnect between what is being presented in the headline versus what is actually present in the article.

iii. **Misleading Conclusion (MLC):** Headline that vaguely draws a conclusion about the article that is not supported by the facts in the article.

Article: అవసరం ఉన్నా లేకపోయినా హీరోయిన్ పాత్ర కు ఒక అక్కనో చెల్లినో పెట్టటం డైరెక్టర్ త్రివిక్రమ్ కి ఉన్న అలవాటు. ఒకరకంగా త్రివిక్రమ్ ఫాలో అయ్యే సెంటిమెంట్లలో ఇది కూడా ఒకటి అని చెప్పవచ్చు. జల్సా, అత్తారింటికి దారేది, అరవింద సమేత సినిమాలలో త్రివిక్రమ్ అదే సెంటిమెంట్ ని ఉపయోగించారు. ఆ సినిమాలు బ్లాక్ బస్టర్ లు అయ్యాయి. అయితే తాజా సమాచారం ప్రకారం త్రివిక్రమ్ తన తదుపరి సినిమాలో కూడా అదే సెంటిమెంట్ ని వాడబోతున్నట్లు వార్తలు వినిపిస్తున్నాయి. మహేష్ బాబు హీరోగా త్రివిక్రమ్ ఒక సినిమా చేయబోతున్న సంగతి తెలిసిందే. ఈ సినిమాలో పూజా హెగ్డే హీరోయిన్ గా నటిస్తోంది. అయితే తాజా సమాచారం ప్రకారం ఈ సినిమాలో సంయుక్త మీనన్ పూజాహెగ్డే సోదరిగా కనిపించబోతున్నట్లు తెలుస్తోంది. త్రివిక్రమ్ స్క్రిప్ట్ అందించిన "భీమ్లా నాయక్" సినిమాలో సంయుక్త మీనన్ రానా భార్య పాత్రలో కనిపించనుంది. ఈ సినిమాలో తన నటనకు ఫిదా అయిన త్రివిక్రమ్ ఆమెను మహేష్ బాబు సినిమాలో కూడా ఎంపిక చేసినట్లు తెలుస్తోంది.

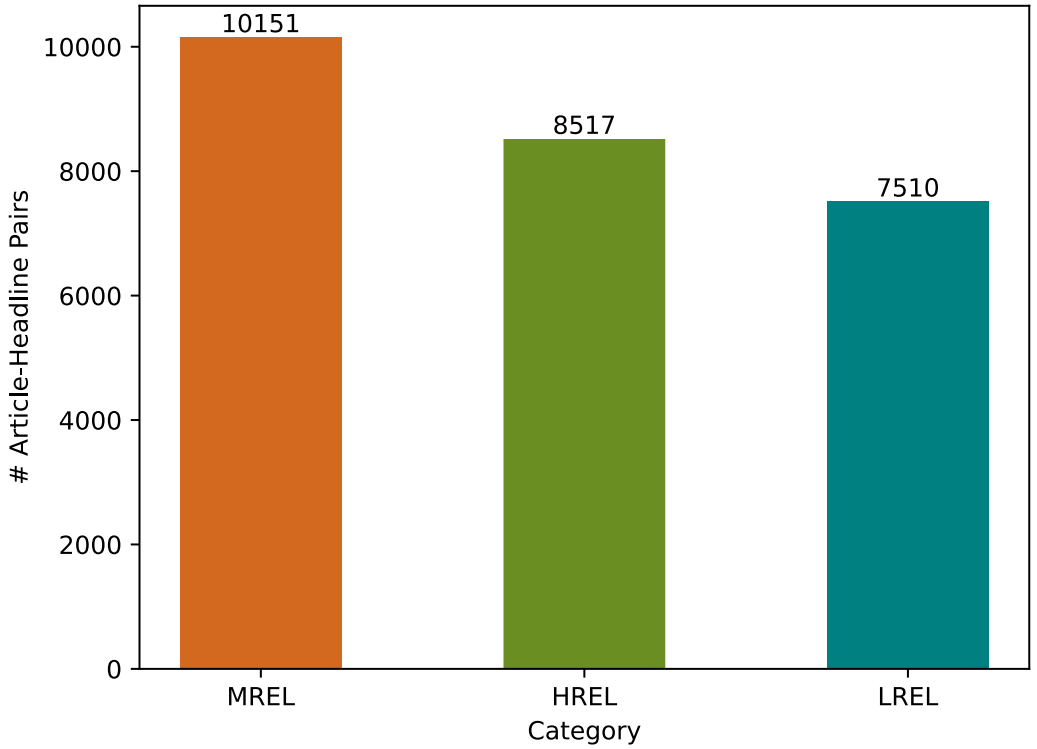
Translation: Director Trivikram's habit is to put an elder sister or sister to the heroine whether it is necessary or not. In a way, this is one of the sentiments that Trivikram follows. Trivikram used the same sentiment in films like Jalsa, Attarintiki Daredi and Aravinda Sametha. Those films became blockbusters. However, according to the latest reports, Trivikram is going to use the same sentiment in his next film as well. It is known that Trivikram is going to do a film with Mahesh Babu in the lead role. Pooja Hegde is playing the female lead in the film. According to the latest reports, Samyuktha Menon will be seen as Pooja Hegde's sister in the film. Samyuktha Menon will be seen essaying the role of Rana's wife in "Bheemla Nayak", which is scripted by Trivikram. Apparently, Trivikram, who was impressed by her performance in the film, has also roped in her for Mahesh Babu's film.

Headline: మహేష్ బాబు సినిమాలో హీరోయిన్ గా రానా వైఫ్

Translation: Rana's wife as heroine in Mahesh Babu's film

Explanation: The article says “Samyuktha Menon (who acted as Rana's wife in Bheemla Nayak movie) to act along with Mahesh Babu in a movie directed by Trivikram” . However, the headline says “Rana's wife as heroine in Mahesh Babu's movie” which is misleading because it deviates from the core information present in the article.

Dataset statistics



- We employed crowd-sourcing for the annotation process
- We assign each article-headline pair to 3 annotators and do majority voting to get final category
- We report an Inter Annotator Agreement score of 0.77, computed using Fleiss Kappa

Dataset statistics

	Train	Dev	Test
# Article-Headline pairs	18324	3927	3927
Avg sentences in article	10.30	10.25	10.29
Avg sentences in headline	1.06	1.06	1.05
Avg words in article	126.33	126.70	126.39
Avg words in headline	6.16	6.15	6.11
(Min, Max) words in article	(27, 226)	(34, 214)	(35, 219)
(Min, Max) words in headline	(2, 22)	(2, 24)	(2, 20)

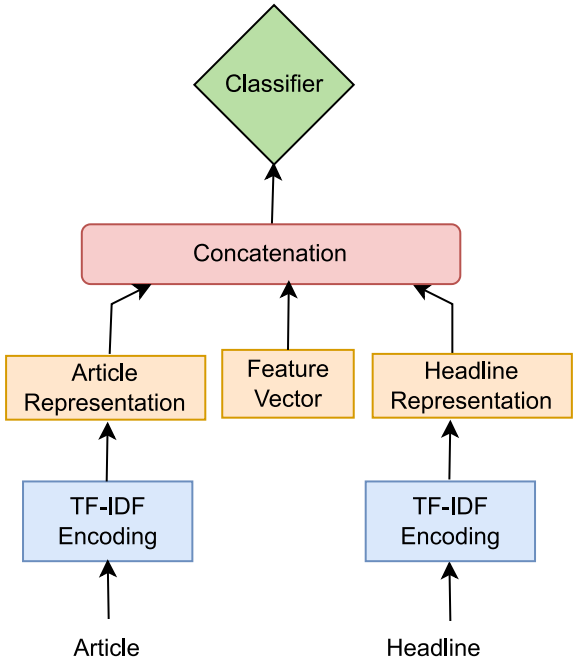
"TeClass" Dataset Statistics

	Train	Dev	Test
HREL	5962	1277	1278
MREL	7105	1523	1523
LREL	5257	1127	1126

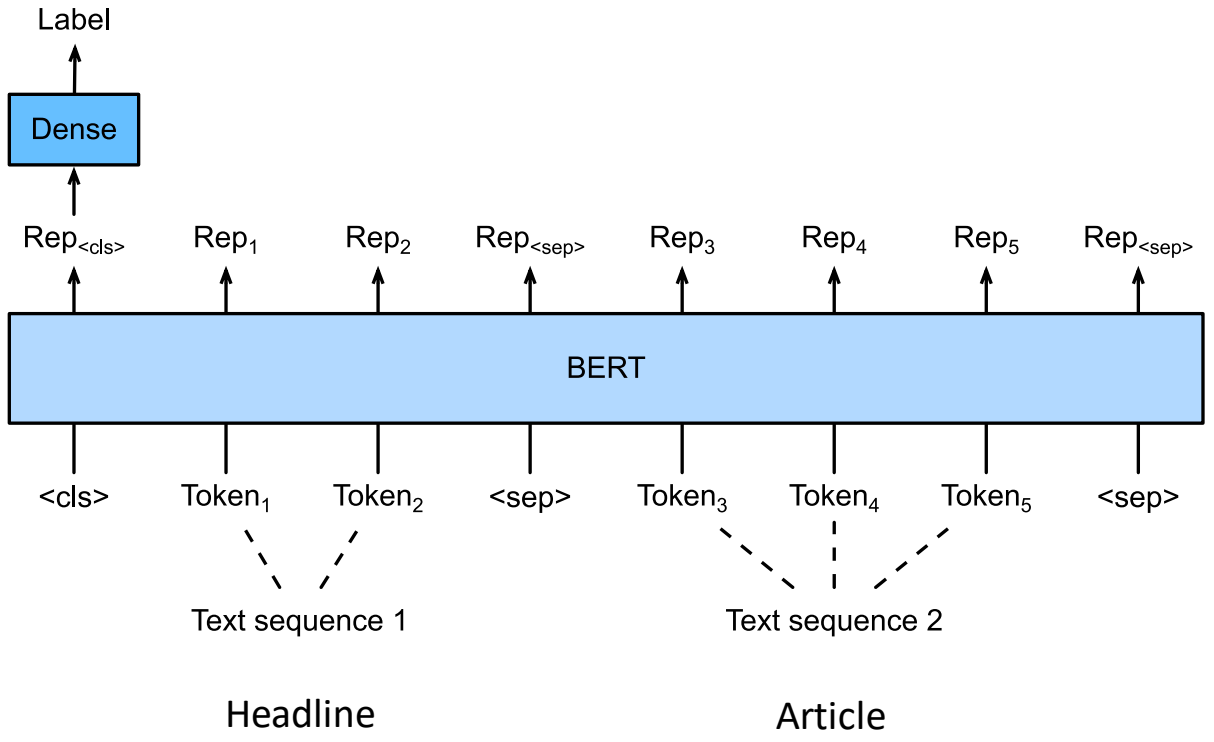
Class-wise distribution in each split

Headline classification baselines

Feature-based ML model



BERT-based model



Headline classification baselines results

Feature-based ML models

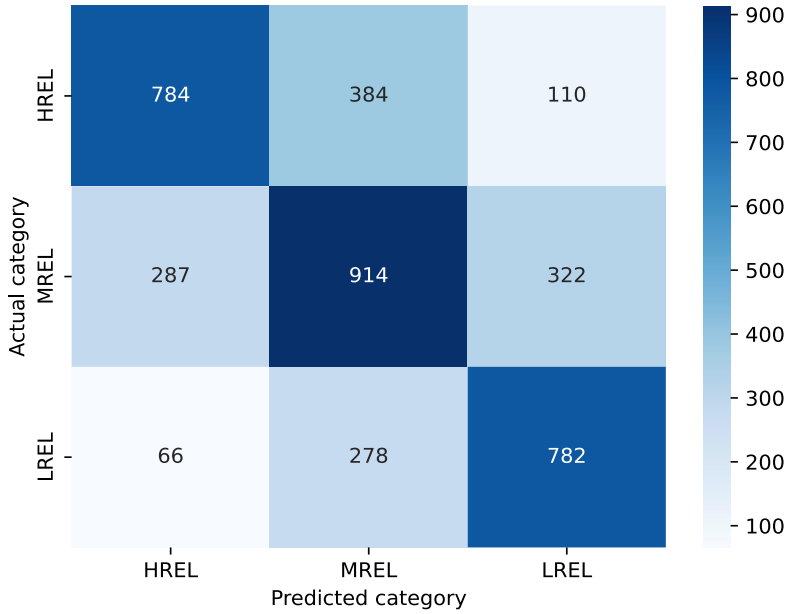
Feature Vector	Classifier	F1 Score				
		HREL	MREL	LREL	Overall (Weighted)	Overall (Macro)
Without Feature Vector	LR	0.57	0.50	0.59	0.55	0.55
	SVM	0.55	0.49	0.57	0.53	0.54
	MLP	0.55	0.49	0.58	0.54	0.54
	Bagging	0.55	0.47	0.57	0.52	0.53
Cosine Similarity	LR	0.58	0.50	0.59	0.55	0.56
	SVM	0.56	0.49	0.58	0.54	0.54
	MLP	0.56	0.49	0.56	0.53	0.54
	Bagging	0.56	0.47	0.58	0.53	0.54
[Cosine Similarity, LEAD-1, Novel 1-gram %]	LR	0.61	0.53	0.59	**0.58**	**0.58**
	SVM	0.60	0.52	0.58	0.57	0.57
	MLP	0.60	**0.54**	0.55	0.56	0.56
	Bagging	0.60	0.51	0.59	0.56	0.57
[Cosine Similarity, LEAD-1, EXT-ORACLE Novel 1-gram %, Novel 2-gram %]	LR	**0.62**	0.53	0.59	**0.58**	**0.58**
	SVM	0.60	0.52	0.58	0.57	0.57
	MLP	0.60	0.50	**0.61**	0.56	0.57
	Bagging	0.60	0.51	0.58	0.56	0.56

Observation: Using Feature Vector along with TF-IDF encoding resulted in better scores compared to the models that do not use feature vector

Headline classification baselines results

BERT-based models

Pre-trained Model	F1 Score				
	HREL	MREL	LREL	Overall (Weighted)	Overall (Macro)
IndicBERT	0.66	0.55	**0.67**	0.62	0.63
mBERT	0.66	0.50	0.62	0.59	0.59
mDeBERTa	0.65	**0.59**	**0.67**	**0.63**	**0.64**
MuRIL	0.66	0.55	0.62	0.61	0.61
XLMRoBERTa	**0.67**	0.53	0.65	0.61	0.62



Observation: BERT-based models performed better than feature-based ML models with mDeBERTa as the best model with a Marco F1 Score of 0.64

- The number of misclassifications between the Highly Related (HREL) and Moderately Related (MREL) classes highlights a notable difficulty: our model struggles to effectively distinguish between these classes
- To improve the performance, we experiment by framing the task as a 2-class classification problem.

Headline classification baselines results

2-Class Classification

Pre-trained model	F1 Score			
	Relevant (FME + STC + FSE)	Less Relevant (WKC+MLC+SEN+CBT)	Overall Weighted	Overall Macro
IndicBERT	0.86	0.66	0.79	0.76
mBERT	0.86	0.63	0.78	0.74
mDeBERTa	0.85	0.69	**0.80**	**0.77**
MuRIL	0.73	0.63	0.70	0.68
XLMRoBERTa	0.86	0.68	**0.80**	**0.77**

Observation: Significantly better performance for BERT-based models with the mDeBERTa model achieving an overall F-1 weighted, macro score of 0.8, and 0.77 respectively.

Relevance-based Headline generation

- We experimented with a pre-trained mT5 headline generation model ”Mukhyansh”, which was trained on a huge corpus of around 825k Telugu article-headline pairs.
- This model was further fine-tuned on different sub-sets of ”TeClass” dataset to evaluate the impact of class-specific fine-tuning on the headline generation task.

Zero-shot Inference of "Mukhyansh" model on "TeClass" dataset

Fine-tune "Mukhyansh" model on class-wise data of "TeClass"

- Only on FME class data
- Only on STC class data
- Only on FSE class data
- Only on WKC class data
- Only on SEN class data
- Only on CBT class data

Fine-tune "Mukhyansh" model on combinations of "TeClass" data

- More relevant classes (FME, STC, FSE)
- Less relevant classes (WKC, SEN, CBT)
- Total 6-class (FME, STC, FSE, WKC, SEN, CBT)

Relevance-based Headline generation results

Fine-tuned on	Tested on						Data Size	
	FME	STC	FSE	WKC	SEN	CBT	Train	Dev
No fine-tuning	0.39	0.23	0.25	0.17	0.21	0.15	-	-
FME	**0.45**	**0.28**	**0.31**	0.21	0.25	0.17	8058	1007
STC	0.43	0.27	0.30	0.22	0.23	0.18	3949	494
FSE	0.41	0.26	0.29	0.22	0.23	0.18	1416	177
WKC	0.38	0.23	0.28	0.20	0.21	0.15	1029	129
SEN	0.41	0.26	0.29	0.20	0.23	0.18	2587	323
CBT	0.39	0.24	0.27	0.21	0.22	0.16	1501	188
Total (6-class)	0.43	0.27	0.30	0.22	0.25	0.18	18540	2318
3-class(FME,STC,FSE)	0.44	**0.28**	0.30	0.20	0.25	0.20	13423	1678
3-class(WKC,SEN,CBT)	0.40	0.25	0.29	0.19	0.23	0.18	5117	640

Observations:

1. No fine-tuning (Zero-shot inference) model performs well enough but if we want the most relevant headline generation then class-aware training always significantly improves (5 points) ROUGE-L score across the board.
2. It is interesting to note that the best performance on all the relevant classes (FME, STC, FSE) is achieved by fine-tuning either on FME class (only 43% of total data) or the combination of all the 3 relevant classes.

Conclusion

Headline Classification

- Presented "TeClass", the first of its kind dataset for relevance-based headline classification
- Explored and experimented with various baseline models including feature-basesd ML, and BERT-based pre-trained models.

Headline Generation

- By experimentation on various sub-sets of "TeClass" dataset,
  - We prove that the task of headline generation is best served when the models are fine-tuned only on highly relevant article-headline pairs compared to fine-tuning on a large dataset having mix of relevant and irrelevant headlines
  - This approach significantly improves the performance and also saves the compute resources

Thank You