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, KCIS, IIIT Hyderabad, India.
{gopichand.kanumolu, lokesh.madasu, nirmal.surange}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

News headline generation is a crucial task in increasing productivity for both the readers and producers of news. This task can easily be aided by automated News headline-generation models. However, the presence of irrelevant headlines in scraped news articles results in sub-optimal performance of generation models. We propose that relevance-based headline classification can greatly aid the task of generating relevant headlines. Relevance-based headline classification involves categorizing news headlines based on their relevance to the corresponding news articles. While this task is well-established in English, it remains under-explored in low-resource languages like Telugu due to a lack of annotated data. To address this gap, we present TeClass, the first-ever human-annotated Telugu news headline classification dataset, containing 78,534 annotations across 26,178 article-headline pairs. We experiment with various baseline models and provide a comprehensive analysis of their results. We further demonstrate the impact of this work by fine-tuning various headline generation models using TeClass dataset. The headlines generated by the models fine-tuned on highly relevant article-headline pairs, showed about a 5 point increment in the ROUGE-L scores. To encourage future research, the annotated dataset as well as the annotation guidelines will be made publicly available.

Keywords: Headline Classification, Headline Generation, Telugu Dataset

1. Introduction

A headline is a single-sentence summary of a news article that aspires to present a concise and factual account of the story described in the article. It is a crucial element in drawing the reader's attention to the article's content and is designed to engage the reader. Headlines are often the only thing that the reader sees before deciding whether to click and read further. They act as a filter, allowing the reader to quickly decide if the story is relevant or interesting to them. In today's rapidly evolving information landscape, the task of assessing the relationship between news headlines and their corresponding articles has become a critical challenge, and this task can be conceptualized in various forms such as fake news detection, misinformation detection, incongruent news headline detection, headline classification, etc.

Generation of a relevant headline can be a challenging and time-consuming task. In most cases, barring sensational and click-bait headlines, the headline needs to draw out the most relevant aspects of the article in a single meaningful string[1]. Therefore, headline generation is often posed as a summarization task (Rush et al., 2015; Gu et al., 2020; Bukhtiyarov and Gusev, 2020). But, despite the existence of multiple article-headline datasets, the generation of relevant headlines remains a challenge, especially for low-resource languages. This can be attributed to the noise present in the datasets in the form of irrelevant headlines (Jin et al., 2020).

The relevance or irrelevance of a headline with respect to the article has been explored by Pomerleau and Rao (2017) in the Fake News Challenge (FNC-1) to determine the stance of a news article relative to the headline. FNC-1 dataset is an extension of the work of Ferreira and Vlachos (2016). The FNC-1 dataset contains 49,972 article-headline pairs labeled with one of the four categories namely Agrees, Disagrees, Discusses, and Unrelated. However, it is important to note that the Unrelated category, constituting 73% of the dataset is generated by pairing the headlines and articles belonging to different topics at random, and hence may not reflect the original relation between article and headline (Chesney et al., 2017).

We believe that the generation of relevant headlines is contingent on the quality of the data presented, especially for low-resource languages like Telugu. We have observed that for low resource languages like Telugu, the ratio of highly relevant headlines versus not-so-relevant or irrelevant headlines is badly skewed towards irrelevance (Figure 1). This might be due to market pressures for publication houses to draw customers to click-baits or might also be due to the cognitively challenging nature of headline creation task. The

*Authors contributed equally
[1]Headline need not be a complete sentence

LREC-COLING 2024, pages 15711–15720

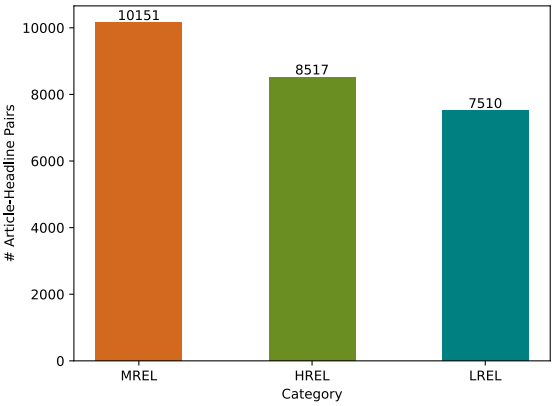

Figure 1: Category distribution in TeClass. HREL: High Relevance, MREL: Medium Relevance, LREL: Low Relevance

impact of this imbalance is seen in wasted time for viewers. Automatic headline generation might help in the latter case but the skew in the distribution of informative headlines means that most of the training compute for the models is spent training on non-informative/irrelevant headlines, eventually impacting the performance negatively. Therefore, we propose that headline generation models should only be trained on highly related article-headline pairs. This requires a pre-processing step of headline relevance classification.

With this motivation, we have created a novel dataset for relevance-based headline classification that reflects the nuances of the real-world news article-headline pairs in the Telugu language. Our key contributions in this paper are summarized as follows:

1. We present "TeClass", a large, diverse, and high-quality human-annotated dataset for a low-resource language Telugu, containing 26,178 article-headline pairs annotated for headline classification with one of the three categories:
   - Highly Related (HREL): The headline is highly related to the article.
   - Moderately Related (MREL): The headline is moderately related to the article.
   - Least Related (LREL): The headline is vaguely related to the article.
2. We present a comprehensive analysis of various baseline models employed for headline classification on this dataset.
3. We present baseline headline generation models to demonstrate that the task of relevant headline generation is best served when the generation models are trained on high-quality relevant data even if the available relevant article-headline pairs are significantly less in number.

To lay the foundation for future work, our dataset and models are made publicly available[2].

2. Dataset

2.1. Selecting the Article-Headline Pairs for Annotation

We collect the news article-headline pairs from multiple news websites using web scraping. As websites often follow their own style of writing the news, to mitigate any potential bias towards a particular style of news reporting, we gathered data from a diverse range of news websites. These websites covered a broad spectrum of domains, including State, National, International, Entertainment, Sports, Business, Politics, Crime, and COVID-19.

However, web scraping from multiple sources posed a significant challenge due to the dynamic nature of websites. Each website has its unique structure, necessitating a thorough understanding of its individual layouts to ensure the extraction of data without loss of information or the introduction of extraneous noise. To address this challenge, we developed custom site-specific web scrapers tailored to each news website. These scrapers were designed to extract three essential components: the text of the news article, the headline, and the name of the news domain. Our extraction methodology was carefully crafted to exclude any undesirable elements, such as advertisements, URLs pointing to related articles, and embedded social media content within the news body.

2.2. Annotation

The relationship between a news headline and its corresponding article can occur in many ways. In ideal cases, the headline summarizes the core idea of the article. Some headlines are designed to capture attention and generate clicks, often by using provocative or sensational language. In some instances, headlines can be misleading, either intentionally or unintentionally, by not accurately representing the information presented in the article. Occasionally, headlines may focus on less important details of the article.

[2] https://github.com/ltrc/TeClass

We employed crowd-sourcing for the annotation process, engaging native Telugu-speaking volunteers. We presented the following instructions to the annotators, and the annotators were asked to assign one of the three primary categories: High relevance (HREL), Medium relevance (MREL), and Low relevance (LREL) after reading the headline and its corresponding article. They are also instructed to assign a secondary sub-class for each article.

HREL: The headline is highly related to the article content if it satisfies the following condition (Example 1 of Figure 2):

- 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.

MREL: The headline is moderately related to the article content if it satisfies any of the following conditions (Example 2 of Figure 2):

- Strong Conclusion (STC): The headline is not explicitly present (in the same words) in the article, but it can be inferred from the article and represents the majority of the article content.
- Factual Secondary Event (FSE): The headline represents a secondary event addressed in the article which is factually correct.
- 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 content.

LREL: The headline is least related to the article content if it satisfies any of the following conditions (Example 3 of Figure 2):

- Sensational (SEN): The Headline is intended to catch the attention of the reader, by reporting biased/emotionally loaded impressions/controversial statements that manipulate the truth of the story.
- Clickbait (CBT): A headline that tempts the reader to click on the link, where there is an extreme disconnect between what is being presented on the front side of the link (headline) versus what is on the click-through side of the link (article).
- Misleading Conclusion (MLC): A headline that vaguely draws a conclusion about the article that is not supported by the facts in the article.
- Unsupported Opinion (USO): A headline that is an opinion about an article's event/subject but is not supported by the article.

A pilot study involving a small-scale trial annotation was conducted to ensure that the annotation guidelines were clear and unambiguous. We explained the guidelines to the annotators to ensure that the annotators understood the task's objectives. Additionally, we closely monitor the annotation process and conduct query resolution sessions to provide assistance in handling ambiguous, or difficult examples. we assign each article-headline pair to 3 annotators, and the final category for a pair is chosen based on the majority vote among the 3 annotations.

2.3. Annotated Dataset Statistics

In this section, we present the statistics of the annotated dataset. Since each article-headline pair is annotated by 3 annotators, we get a total of 78,534 annotations for 26,178 unique article-headline pairs. The category-wise counts of the dataset are presented in Figure 1. As mentioned earlier, the dataset contains article-headline pairs from multiple websites with a diverse set of news domains, the website-wise and domain-wise pairs distribution is detailed in Figure 3, and Figure 4 respectively.

Data Splits: We allocated 70% for training, 15% for development and 15% for testing. To ensure unbiased performance and prevent category bias, we applied stratified sampling techniques. This ensures even distribution of articles from all 3 categories across the training, development, and test sets. The category-wise counts in each data split are presented in Table 1. Further statistical details of the TeClass dataset are available in Table 2.

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

Table 1: Category-wise counts in each data split

Inter-Annotator Agreement: Having multiple annotators (typically three or more) for annotation tasks is vital for several reasons. They enable the measurement of inter-annotator agreement, helping to identify and address ambiguous or challenging cases. Multiple annotators also help mitigate individual bias and promote a balanced, objective annotation process ensuring the robustness and quality of the annotated dataset. We use Fleiss' Kappa metric proposed by Randolph (2005) and it resulted in an encouragingly high score of 0.77, indicating a substantial agreement among the annotators.

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

Category: Highly Related

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.

Example 1: Highly Related Headline

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

Category: Moderately Related

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."

Example 2: Moderately Related Headline

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 స్క్రిప్ట్లో 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

Category: Least Related

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.

Example 3: Least Related Headline

Figure 2: Examples of relevance-based headline classification for each category

	Train	Dev	Test
Article-Headline pairs	18,324	3,927	3,927
Average sentences in article	10.30	10.25	10.29
Average sentences in headline	1.06	1.06	1.05
Average tokens in article	126.33	126.70	126.39
Average tokens in headline	6.16	6.15	6.11
Unique tokens in articles	204959	76279	76070
Unique tokens in headlines	28785	9894	10008
Average LEAD-1 score	16.88	17.09	16.88
Average EXT-ORACLE score	29.47	29.01	29.49

Table 2: TeClass Statistics

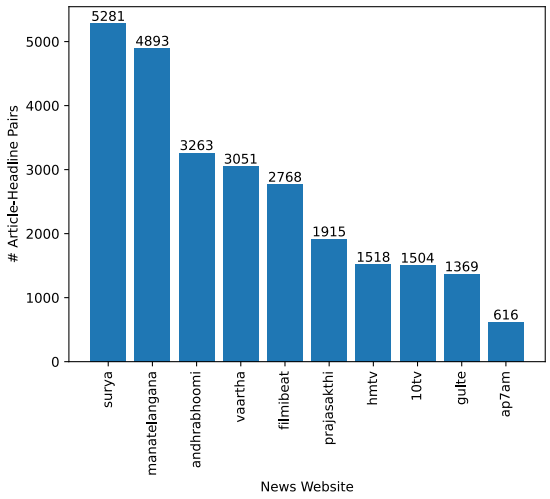

Figure 3: News website distribution in TeClass

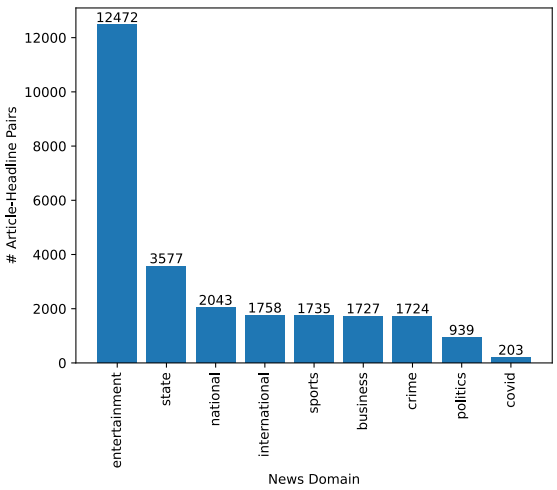

Figure 4: News domain distribution in TeClass

3. Headline Classification

We experiment with various baseline models including traditional feature-based Machine Learning (ML) models for classification, and also leverage the transfer learning using the state-of-the-art pre-trained BERT (Devlin et al., 2018) models.

ML baseline models: Various participating teams in the FNC-1 challenge make use of features like n-gram overlap, cosine similarity between vector representations of the article, and the headline, and other hand-crafted features (Hanselowski et al., 2018). We also experiment with various features, and our model architecture is similar to the one proposed by Riedel et al. (2017) . We use TF-IDF encoding to represent the article, and headline in vector format. To avoid the problem of out-of-vocabulary words, we use subword tokenization that breaks words into smaller subword units, which is vital for morphologically rich languages like Telugu. It resulted in a subword vocabulary of size 2945, which is in turn the dimension of the vector representation of the article, and headline using TF-IDF encoding. We concatenate the feature vector with the article, and headline representations, and the output of concatenation is passed as input to train the classifier. The feature vector is extracted from the article-headline pairs using the following methods:

1. Cosine similarity: To measure the similarity in content between the article and headline, we compute the cosine similarity between the TF-IDF vector representations of the article and headline.

2. Novel n-gram percentage: It quantifies the level of uniqueness in a headline by measuring the proportion of n-grams (contiguous sequences of n words) found in the headline but not present in the accompanying article.

3. LEAD-1: It is the ROUGE-L (Hasan et al., 2021)[3] score between the headline and the first sentence of the article.

[3] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

4. EXT-ORACLE: This score is computed by selecting the sentence from the article that achieves the highest ROUGE-L score with the headline.

We use Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Bagging as classification models. All these models use 5-fold cross-validation. We assess model performance using the F1-Score, and the corresponding results are presented in Table 3.

BERT-based baseline models: Pre-trained models like BERT excel in text classification compared to classical ML models because they leverage extensive pre-training on diverse data, capturing language nuances and context. In our work, we fine-tuned several state-of-the-art multilingual BERT-based models, equipping them with a classification head. The classification head is a feedforward neural network added on top of the BERT model, specifically trained for our classification task. We used a specific input format where the headline and news article text were concatenated, separated by a [SEP] token, and preceded by a [CLS] token. This format ensures a unified representation of both the title and text, significantly enhancing the model's ability to process and make accurate predictions.

We experiment with the following models by making use of the scripts [4] provided by Huggingface.

mBERT: mBERT (Devlin et al., 2018) is a multilingual variant of the BERT model, which supports 102 different languages. For our baseline, we fine-tune the base version of mBERT having 110M parameters.

XLM-RoBERTa: XLM-RoBERTa (Conneau et al., 2019) is a multilingual version of the RoBERTa model, and it was pre-trained on a vast 2.5TB CommonCrawl dataset, which included text from 100 languages. For our experiments, we utilized the xlm-roberta-base variant, boasting 270 million parameters.

MuRIL: MuRIL (Khanuja et al., 2021) is pre-trained on 17 Indian languages, utilizing a range of datasets, including Wikipedia, CommonCrawl, PMINDIA, and Dakshina Corpora. We employed the muril-base-cased variant with 236 million parameters for our task.

IndicBERT: IndicBERT (Doddapaneni et al., 2023) is a multilingual BERT model trained with the Masked Language Modeling (MLM) objective on the IndicCorp v2 dataset. This model supports 23 Indic languages as well as English and boasts 278 million parameters. We used the IndicBERTv2-MLM-only version in our experiments.

mDeBERTaV3: mDeBERTaV3 (He et al., 2021) is a multilingual adaptation of the DeBERTa model, pre-trained on a substantial 2.5TB dataset known as CC100, featuring text from 100 languages. We used the base variant of mDeBERTaV3 in our experiments.

Hyperparameters: For all these models, we set the maximum input sequence length to 512 subword tokens, and use a batch size of 8. We use categorical cross-entropy loss with Adam optimizer and a learning rate of 2e-05. To prevent overfitting, we use early stopping criteria to stop training when the validation loss stops improving (or begins to worsen) over two consecutive epochs. All these experiments were performed using 4 GPUs (each with a VRAM of 12GB), and 30 CPUs. The results of these experiments are presented in Table 4.

4. Results & Analysis

From the results presented in Table 3, it is apparent that the integration of a feature vector in conjunction with TF-IDF encoding, featuring elements such as cosine similarity, LEAD-1, EXT-ORACLE, Novel 1-gram %, and 2-gram %, clearly underscores the vital role played by these features in enhancing the performance of our models when compared to models that did not employ a feature vector. Notably, the Logistic Regression (LR) model utilizing these features achieved F1 weighted and macro scores of 0.58, which represents a 3% improvement when compared to the model that did not utilize a feature vector.

Furthermore, the results presented in Table 4 underscore the superiority of state-of-the-art BERT-based models in comparison to classical machine learning models. The best model, mDeBERTa, achieved an impressive overall F1 weighted score of 0.63 and an F1 macro score of 0.64. These scores reflect a substantial 5% improvement in F1 weighted and a 6% improvement in F1 macro scores when compared to the best-performing feature-based ML model.

The confusion matrix between actual categories and predicted categories of the mDeBERTa model shown in Figure 5 offers valuable insights into the challenges encountered by our model. Specifically, 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

[4] https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification

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

Table 3: Headline Classification: ML baseline model results

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

Table 4: Headline Classification: BERT baseline model results

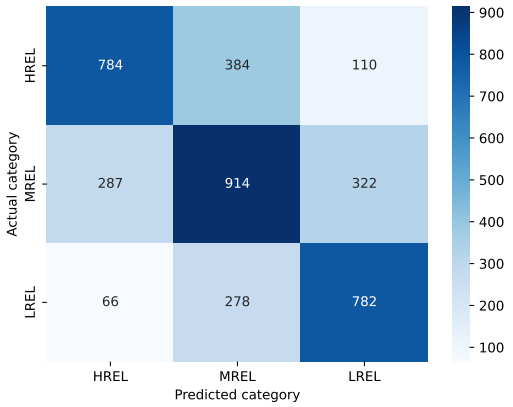

Figure 5: Confusion matrix between actual and predicted categories of mDeBERTa model

these classes. But, if we consider Factual Main Event, Factual Secondary Event and Strong Conclusion classes as relevant to the article, we see significantly better performance for DL models as seen in Table 5. This underscores the inherent difficulty in differentiation between highly relevant and moderately related headlines.

5. Headline Generation

We experimented with headline generation by using mT5 model trained on Telugu summary generation on a large Telugu dataset (Mukhyansh (Madasu et al., 2023)). This was further fine-tuned on different subsets of TeClass to evaluate the impact of class-specific fine-tuning on the headline generation task. As seen in Table 6, non-fine-tuned 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. In a human evaluation conducted by two volunteers on 50 news articles, we found that 34, 1, and 3 generated headlines were marked as FME, FSE, and STC respectively.

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 or the combination of all the relevant classes. It is also interesting to see that the performance gain is not

Pre-trained model	F1 Score			
	FME+FSE+STC	SEN+WKC+USO+MLC+CBT	Overall(Weighted)	Overall(Macro)
IndicBERT	0.86	0.66	0.79	0.76
mBERT	0.85	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**

Table 5: Headline Classification: BERT baseline model results for Merged fine classes

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

Table 6: Class-based Headline Generation results. (Metric: ROUGE-L)

proportional to the training data size. In fact, we see a marked decrease in performance when all of the data is used. The best performance is achieved using 43% of the data (FME).

6. Conclusion & Future work

In this work, we introduce a novel, high-quality human-annotated dataset tailored for the task of relevance-based news headline classification in a low-resource language, Telugu. Our proposed dataset comprises 26,178 article-headline pairs, meticulously annotated into three primary classes: Highly Related, Moderately Related, and Least Related. Notably, this dataset stands as the largest and most diverse of its kind, encompassing various news domains and websites. This contribution marks the first dataset of its nature specifically designed for the task of headline classification in the Telugu language.

In our experiments with various baseline models on this dataset, our empirical findings highlight the superior performance of BERT-based models when compared to classical machine learning models. Notably, mDeBERTa achieved an impressive F1 weighted score of 0.63 and an F1 macro score of 0.64. We firmly believe that this dataset will serve as a valuable resource for the research community working on applications such as News Headline Classification, Fake News Classification, Misinformation Classification, and other related tasks. Furthermore, the annotation guidelines and annotation process developed for this dataset can be a valuable reference for extending this task to other languages.

Further, this classification of these headlines into relevance classes assists significantly in generation of high-quality headlines at half the compute cost (with respect to a number of samples). We hope that this work will encourage attempts to extract high-quality data for generation tasks in general.

7. Ethics Statement

The collected news articles are subject to the respective licenses of the original websites. These resources will be released under the Creative Commons license[5], respecting individual website policies on data distribution and public availability.

8. Acknowledgments

We extend our sincere gratitude to Pavan Baswani for generously providing the annotation tool, which facilitated the acquisition of high-quality annotations.

9. Bibliographical References

[5] https://creativecommons.org/licenses/by/4.0/

Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of transformer-based models for news headline generation. In *Artificial Intelligence and Natural Language*, pages 54–61, Cham. Springer International Publishing.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: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.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1163–1168. The Association for Computational Linguistics.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5082–5093. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Lokesh Madasu, Gopichand Kanumolu, Nirmal Surange, and Manish Shrivastava. 2023. Mukhyansh: A headline generation dataset for Indic languages. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 620–634, Hong Kong, China. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on*

Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.