A Benchmark for Relevance-based Headline Classification and Generation

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Gopichand Kanumolu
2021701039
gopichand.kanumolu@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, India
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**A Benchmark for Relevance-based Headline Classification and Generation**" by Gopichand Kanumolu, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Manish Shrivastava

This Research Thesis is Dedicated to My Family and Friends

Acknowledgments

I extend my heartfelt gratitude to all those who have supported and guided me throughout this research journey. First and foremost, I am deeply thankful to my advisor, Dr. Manish Shrivastava, for their invaluable mentorship, support, and encouragement. Their expertise, insights, and guidance have been instrumental in shaping this thesis and my overall academic growth.

I would like to express my heartfelt gratitude to my parents, Satyanarayana Kanumolu and Vijayalakshmi Kanumolu, for their boundless love, understanding, and support. Their belief in me has been a constant source of motivation and strength. I am thankful to my brother, Venkat Kanumolu, for his strong support; he has been my pillar of strength throughout my academic journey. I am grateful to my friends and collaborators Lokesh Madasu and Pavan Baswani for their companionship, camaraderie, and encouragement. Their presence has made this academic pursuit a more enjoyable and fulfilling experience.

Thankful to Ashok Urlana for his help in beginning my research internship journey. Special thanks to my senior PhD students (Nirmal, Ananya, Prashanth, Hiranmai, Aparajita) and MS students (Ganesh, Manikanta, Rakesh) of LTRC (MT-NLP lab) for their guidance, and willingness to share their knowledge. Lastly, I would like to acknowledge the support and contributions of all individuals from various RGUKTs who have directly or indirectly helped in the completion of this thesis. Your assistance has been invaluable and deeply appreciated.

Thank you all for being part of this journey.

Abstract

Keywords: Headline Classification, Headline Generation, Telugu Dataset.

The task of news headline generation deals with generating a concise summary for a given news article. It is a crucial task in increasing productivity for both the readers and producers of news. Significant progress has been made in automatically generating headlines for widely spoken languages like English. A notable obstacle hindering headline generation in Indian languages is the lack of high-quality data. To address this gap, we present "Mukhyansh", a comprehensive multilingual dataset collected from the web for the task of Indian language headline generation. Mukhyansh contains over 3.39 million article-headline pairs across eight prominent Indian languages: Telugu, Tamil, Kannada, Malayalam, Hindi, Bengali, Marathi, and Gujarati. The news data collected from various websites on the web comprises a mixture of both relevant and irrelevant headlines, including sensational, clickbait, and misleading ones. As a consequence of these irrelevant headlines in scraped news articles, headline generation models often produce sub optimal results. We propose a novel approach centered on relevance-based headline classification to enhance the performance of headline generation models. Relevance-based headline classification deals with categorizing news headlines based on their relevance to corresponding articles. While relevance-based headline classification is well-established in English, its application in low-resource languages like Telugu remains largely unexplored due to a scarcity of annotated data. Our study aims to address this gap by introducing the "TeClass" dataset, the first-ever human-annotated relevance-based Telugu news headline classification dataset. The proposed dataset contains 78,534 annotations across 26,178 article-headline pairs, making it the largest publicly available dataset of its kind. We experiment with various baseline models on this dataset and provide a comprehensive analysis of the model results. The annotated dataset as well as the annotation guidelines, and models are made publicly available to encourage future research.

Furthermore, we utilize this dataset to illustrate the impact of fine-tuning headline generation models on various headline categories, each exhibiting different degrees of relevance to their respective articles. Our empirical results demonstrate that the performance of headline generation models is enhanced when models are fine-tuned on datasets containing highly relevant headlines, despite the smaller quantity of such data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Relevance-based Headline Generation

Generating relevant headlines for news articles is indeed a challenging task, particularly due to the varying quality and nature of the training data. If the data contains a mix of relevant headlines along with clickbait, sensational, or misleading ones, it can lead to a model that learns to generate similarly problematic headlines. Ensuring high-quality, relevant training data is essential to train models that produce accurate and meaningful headlines. Headline generation requires understanding the context and main points of the news article. This can be challenging, especially for complex or nuanced topics where the key information may be buried within the article. Models need to effectively capture the essence of the article to produce relevant headlines.

Headlines should accurately represent the content of the article while also being engaging and attention-grabbing. Striking the right balance between accuracy and creativity is challenging. Models need to generate headlines that are both informative and enticing to readers without resorting to sensationalism or misleading tactics. Headline generation models should be able to adapt to a wide range of news topics and domains. Generalizing across diverse topics while maintaining relevance and accuracy is a significant challenge. Models need to capture the nuances of different subjects and generate headlines that are appropriate for each topic.

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 [38, 14, 2]. 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 [21].

We believe that the generation of relevant headlines is contingent on the quality of the data presented to the models during training, 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

[1]Headline need not be a complete sentence

or irrelevant headlines is badly skewed towards irrelevance (Figure 1.1). This might be due to market pressures for publication houses to draw customers to click-bait or might also be due to the cognitively challenging nature of headline creation task. The 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 relevance-based headline classification, that can greatly aid the task of generating relevant headlines.

1.2 Relevance-based Headline Classification

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. News headlines and articles serve as the initial gateway through which individuals access information. However, in today's digital age, the competition for clicks and views has led to a phenomenon where some media outlets prioritize sensationalism over accuracy. Misinformation and fake news often thrive in this environment, as attention-grabbing headlines and articles generate more traffic and consequently more revenue. This creates a perverse incentive for some outlets to prioritize sensationalism and clickbait over factual reporting. As a result, misinformation can spread rapidly, eroding public trust in media and distorting perceptions of reality.

The spread of misinformation in news or fake news has profound effects on both users/readers and news-making agencies. For users, exposure to false or misleading information can lead to confusion, mistrust in media sources, and polarization of beliefs. Misinformation can also have real-world consequences, influencing decisions ranging from personal health choices to political opinions. News agencies, on the other hand, face significant challenges when misinformation spreads through their platforms. It undermines their credibility and integrity, leading to diminished trust among audiences and potential legal ramifications. Additionally, combating misinformation often requires significant resources and can divert attention from other important journalistic endeavors.

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. Researchers in Natural Language Processing (NLP) are actively working to combat the spread of misinformation in news and fake news through various approaches. One key strategy involves developing

advanced machine learning models capable of detecting and classifying misinformation at scale. NLP researchers are leveraging techniques such as text classification, and semantic understanding to identify misleading or false information in news articles and social media posts. Classification approaches commonly used include supervised machine learning algorithms, deep learning models, and Transformer-based architectures. These models are trained on large datasets of labeled examples, enabling them to automatically categorize content as either trustworthy or deceptive based on linguistic cues, factual inconsistencies, and contextual information. Additionally, researchers are exploring techniques for detecting subtle forms of misinformation, such as framing effects and propaganda, to enhance the effectiveness of detection systems.

The scarcity of human-annotated, high-quality datasets for misinformation/fake news headline classification tasks in low-resource languages like Telugu presents a significant challenge for AI/NLP research. For the task of news headline classification, having access to diverse and well-annotated datasets is crucial for training and evaluating machine learning models effectively. However, for languages with fewer resources, such as Telugu, building such datasets is particularly challenging. These datasets are essential for advancing AI/NLP research for several reasons. Firstly, they serve as the foundation for training robust machine learning models that can accurately detect and classify misinformation in Telugu news headlines. Without sufficient data, models may struggle to generalize well to real-world scenarios, leading to poor performance and unreliable results. In summary, the availability of human-annotated, high-quality datasets for misinformation classification tasks in low-resource languages like Telugu is essential for driving NLP research forward. By addressing the scarcity of such datasets and investing in their creation, researchers can develop more robust and effective misinformation detection systems, ultimately helping to combat the spread of false information and promote digital literacy in diverse linguistic communities.

With this motivation, in this work, we have created a novel, Telugu human-annotated dataset namely "TeClass", for the task of relevance-based headline classification that deals with classifying the headlines based on their relevance to the news article. This dataset, the first of its kind for Telugu, comprises 26,178 article-headline pairs meticulously annotated for headline classification into three categories: Highly Related (HREL), Moderately Related (MREL), and Least Related (LREL). By focusing on the relevance of headlines to news articles, this dataset provides valuable resources for training and evaluating machine learning models tailored to the nuances of Telugu language and news media. Researchers can leverage this dataset to develop more accurate and effective algorithms for Telugu news classification. In addition, this dataset can serve as a valuable resource for studying the impact of different types of news headlines on the performance of headline generation models.

1.3 Key Contributions

Our key contributions in this work are summarized as follows:

1. We introduce "Mukhyansh", a multilingual headline generation dataset containing 3.39 Million news article-headline pairs across 8 Indian languages Telugu, Tamil, Kannada, Malayalam, Hindi, Marathi, Bengali, and Gujarati. We provide various baseline models for the task of headline generation.

2. 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 the task of relevance-based headline classification with one of the three categories: Highly Related, Moderately Related, and Least Related. We provide a comprehensive analysis of various baseline models employed for headline classification on "TeClass" dataset.

3. Further, we use the "TeClass" dataset to study the impact of fine-tuning headline generation models on various categories of headlines with varying degrees of relevance to the article and demonstrate that the task of relevant headline generation is best served when the generation models are fine-tuned on high-quality relevant data even if the available relevant article-headline pairs are significantly less in number.

Our datasets and models are publicly available[2][3] to lay the foundation for future work.

1.4 Thesis Outline

The remaining sections of this work are organized as follows: In Chapter 2, we discuss the existing works in headline generation and headline classification. Chapter 3 provides the details of the "Mukhyansh" dataset, headline generation models for Indian languages, and creation of the "TeClass" dataset, annotation guidelines. Chapter 4 presents the details of relevance-based headline classification baseline models. In Chapter 5, we provide the details of relevance-based Telugu headline generation models and results. Chapter 6 concludes with a summary of our contributions.

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

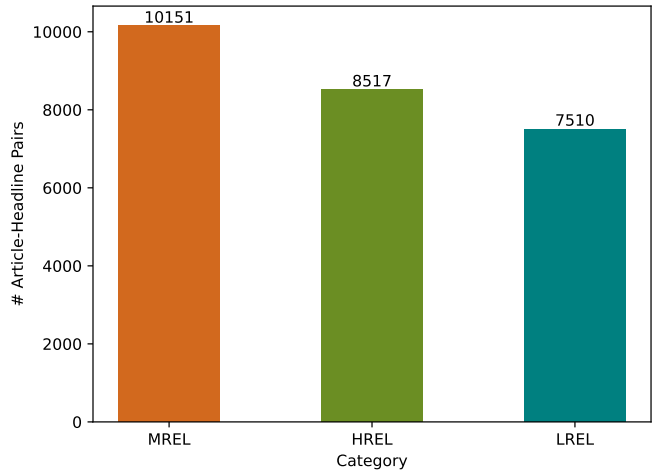

Figure 1.1 Category distribution in TeClass. HREL: Highly Related, MREL: Moderately Related, LREL: Least Related

Chapter 2

Literature Survey

2.1 Headline Generation

Headline generation, a pivotal task in natural language processing, involves condensing the key information from a news article into a single, short sentence. Over time, the NLP community has made significant strides in advancing techniques for this task. Various approaches have been explored, ranging from rule-based methods to deep learning models. These approaches leverage diverse strategies such as extractive summarization, abstractive summarization, or a combination of both.

In the ”parse-and-trim” approach [10], the article is parsed into its constituent parts using a dependency parser, and the most relevant parts are selected to form the headline using linguistically motivated heuristics. A statistical approach is proposed in [1] by considering the problem of headline generation as analogous to statistical machine translation. This approach uses a language model to predict the words in the headline given the words in the source article. The application of neural networks and deep learning for NLG has enabled a major shift in the field, leading to more sophisticated and effective models that can generate high-quality text. An attention mechanism-based summarization (ABS) method was presented in [39], which utilizes a convolutional network in the encoder to encode the input words and uses attention on top of it to help the text generation model. While this approach uses a feed-forward neural network for text generation, a Recurrent Neural Network (RNN) based decoder proposed in [5] improved the performance when compared to the previous approach on the English Gigaword dataset. [34] proposed a novel Encoder-Decoder network with a hierarchical attention mechanism that utilizes recurrent neural networks in both the encoder and decoder side. In [13], COPYNET mechanism that allows the model to generate the text by copying the words from the input is incorporated. This approach is successful in handling the Out-Of-Vocabulary (OOV) words issue. A pointer generator network that generates a probability distribution over vocabulary and can copy words from the input using attention distribution is proposed by [40] which is effective in handling OOV words problem. Furthermore, the authors introduce a novel coverage mechanism to deal with the problem of repetitive words on the decoder side. While the previous approaches aim to generate plain, and factual headlines, a stylistic

headline generation system that is capable of generating headlines with three target styles: humor, romance, and clickbait is developed in [22].

Additionally, the availability of datasets plays a crucial role in shaping the development of headline generation systems. The quality of datasets directly impacts the performance of headline generation models. High-quality datasets provide diverse and representative examples that enable models to learn effectively. They encompass a wide range of topics, writing styles, and linguistic nuances present in news articles. The majority of prior research in this area has been conducted on the resource-rich language English, while exploration of low-resource languages remains relatively understudied.

XL-Sum [17] is a multilingual abstractive summarization dataset that can also be used for headline generation. The Indian language section of XL-Sum dataset contains 251K article-headline pairs collected from a single source BBC[1]. To advance further research in NLG for Indian languages, [26] proposed IndicNLG benchmark consisting of datasets for five different NLG tasks, which includes headline generation. For the task of headline generation, they released a multilingual dataset (henceforth referred to as IndicHG dataset) consisting of 1.31 million article-headline pairs spanning 11 Indian languages and implemented baseline models by fine-tuning the pre-trained encoder-decoder transformer models IndicBART [7] and mT5-small [46].

However, in our analysis, detailed in Chapter 5, we demonstrate that the IndicHG dataset suffers from serious quality issues like data contamination, and data duplication and could not be considered ideal for training robust NLG models for the task of headline generation. To overcome these issues, we present a massive multilingual headline generation dataset named 'Mukhyansh' [29], containing 3.39 Million article-headline pairs across 8 Indian languages including Telugu, Hindi, Bengali, Marathi, Tamil, Kannada, Malayalam, Gujarati.

While these studies focus on training headline generation models using scraped data collected from the web, they did not specifically examine the impact of relevant or irrelevant headlines in the training data on the models' performance. With this motivation, our research investigates the effect of relevant and irrelevant headlines in the training data on the performance of headline generation models.

2.2 Headline Classification

News headline classification is a subfield of Natural Language Processing (NLP) concerned with automatically categorizing news headlines into different predefined classes or identifying the relationship

[1] https://www.bbc.com/

between headlines and the corresponding articles. This process helps organize news content, personalize user experience, and combat misinformation. This task is approached in different forms:

- **Topic Classification**: Here, the objective is to assign a single category label to a headline. This category can represent the topic of the news (business, sports, entertainment, politics, etc.). This task helps in organizing news articles and facilitating content recommendation systems [23][32].

- **Sentiment Analysis**: It involves determining the sentiment conveyed in a headline, such as positive, negative, or neutral. This task helps in understanding public opinion, brand perception, and market trends [30].

- **Clickbait Classification**: Clickbait headlines are designed to be sensational and misleading, enticing users to click without accurately reflecting the article's content. Classification models can identify keywords, phrasing patterns, and sentiment often associated with clickbait headlines [31].

- **Headline-Article Relation Classification**: This task goes beyond simple categorization. It focuses on understanding the relationship between a headline and its corresponding article. Here, models analyze the headline and the full article content to determine the stance or perspective the headline takes on the reported event (for example, does the headline accurately reflect the article's neutrality, or is it biased?). Stance classification is an important NLP task that deals with determining the relationship between a headline and its associated article. It aims to classify whether the headline supports, contradicts, or remains neutral on the topics discussed in the article. Stance classification plays a pivotal role in combating the spread of misinformation. This task is particularly challenging because of the complex relationships between a headline and an entire news article such as variation in length (number of tokens) and linguistic complexities. This task can be conceptualized in various forms such as fake news detection, misinformation detection, headline stance classification, incongruent news headline detection, headline classification, etc.

There is a considerable amount of research conducted on headline stance classification [25] by the NLP community. Most of these studies are focused on resource-rich languages like English due to the presence of publicly available datasets. One of the most widely used datasets for stance detection includes the English tweets corpus by [33]. It comprises 4,870 tweet-target pairs annotated to determine the tweeter's stance towards six predefined targets. Here, the target can be any entity (person, organization, policy, etc.) or topic (Climate change, Atheism, etc.) and the task is to determine whether the tweeter is in favor, against, or neutral towards the target. A similar kind of work is done in a few Indian languages. The dataset proposed by [43] contains 3,545 English-Hindi code mixed tweets annotated with a stance towards the Demonetisation topic. [41] introduced a corpus containing 1,450 English-Kannada code mixed Facebook posts and comments related to four targets annotated for the task of stance classification.

While previous studies focus on detecting the stance of individual sentences concerning a target topic or entity, the relevance or irrelevance of a headline concerning the article has been explored by [35] in the Fake News Challenge (FNC-1) to determine the stance of a news headline relative to the article. FNC-1 dataset is an extension of the work of [11]. 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 [3].

Existing research on classifying the relationship between a news headline and its corresponding article has primarily focused on high-resource languages like English. There is a lack of research on this topic for languages with fewer resources, such as Telugu. A major challenge in this area is creating a high-quality human-annotated dataset that captures the nuances of real-world news. To address this gap, we present ”TeClass,” the first-ever human-annotated dataset for Relevance-based Headline Classification in Telugu. TeClass is the largest publicly available dataset of this kind, containing a total of 26,718 article-headline pairs annotated for headline classification into one of three categories: Highly Related (HREL), Moderately Related (MREL), and Least Related (LREL).

Chapter 3

Mukhyansh & TeClass Datasets

3.1 Introduction

Headline generation plays a crucial role in summarizing news articles and capturing readers' attention. The task of headline generation involves automatically generating informative and captivating headlines that accurately capture the essence of the underlying text. In recent years, the NLP community has achieved remarkable strides in the development of headline-generation models. However, the focus has primarily been on English and other widely spoken languages, inadvertently leaving a significant void in the space of headline generation for Indian languages. One of the most significant obstacles hindering headline generation in Indian languages is the scarcity of high-quality datasets.

3.1.1 Issues in existing headline generation datasets

In an attempt to fill the gap in the availability of datasets for the task of Indian language headline generation, researchers in [26] proposed the IndicNLG benchmark, encompassing five different NLG tasks, including a headline generation dataset (hereafter referred to as IndicHG dataset). IndicHG dataset comprises 1.31 million article-headline pairs across 11 Indian languages. Surprisingly, our analysis reveals that the IndicHG dataset contains a significant number of duplicate article-headline pairs in the training, development, and test splits for most languages. Out of the total 1.31 million pairs, approximately 0.67 million (51.23%) are duplicates. Moreover, it is ideal for a dataset to have no overlap or common samples among the training, development, and test splits. However, the statistics presented in Table 3.1 demonstrate a high level of data overlap or contamination among these splits for most of the languages. For instance, an article-headline pair[1] from the Kannada language appears 115 times in the training data, 18 times in the development data, and 2 times in the test data.

[1] `https://tinyurl.com/2p85mayt`

Models trained on contaminated data may learn to prioritize certain article-headline pairs that frequently appear in the dataset, biasing them towards generating similar headlines and potentially sacrificing diversity and creativity. Overfitting to duplicated samples can occur, where the model memorizes them instead of learning meaningful patterns, leading to degraded performance on unseen data. Evaluation metrics computed on contaminated test sets may inaccurately reflect the model's true performance. Additionally, these models may struggle with variations in headline generation tasks, particularly when faced with unseen or out-of-distribution data, due to the lack of diverse training examples, hampering their ability to generalize effectively to real-world scenarios.

L	Train set		Development set			Test set			Total		
	# Pairs	Duplicates (%)	# Pairs	Duplicates (%)	Train Overlap (%)	# Pairs	Duplicates (%)	Train-Dev Overlap (%)	# Pairs	(Duplicates + Overlap) (%)	Remaining
te	21352	8.77	2690	1.52	15.61	2675	1.42	18.61	26717	10.38	23945
ta	60650	51.18	7616	50.22	3.31	7688	50.20	3.62	75954	51.29	36996
kn	132380	87.26	19416	84.29	59.18	3261	6.23	71.17	155057	87.51	19364
ml	10358	22.83	5388	76.26	33.33	5220	76.05	44.22	20966	53.78	9690
hi	208091	3.19	44718	0.76	6.42	44475	0.72	7.83	297284	4.59	283646
bn	113424	69.86	14739	68.02	19.41	14568	67.94	24.30	142731	70.65	41896
mr	114000	69.10	14250	66.95	15.45	14340	67.03	16.15	142590	69.73	43157
gu	199972	75.11	31270	80.04	0.96	31215	80.02	1.28	262457	76.33	62123
pa	48441	0.13	6108	0	0.18	6086	0	0.35	60635	0.16	60540
as	29631	30.05	14592	75.96	58.77	14808	75.97	65.91	59031	60.66	23222
or	58225	48.77	7484	48.97	0.16	7137	48.58	0.42	72846	48.79	37305
								Total:	1316268	51.23	641884

Table 3.1 IndicHG Analysis: data duplication and overlap (data-contamination) percentages in training, development and test datasets of IndicHG dataset

3.2 Mukhyansh Dataset

To overcome the issues in existing datasets and to bridge the gap in advancing research for the task of headline generation in Indian languages, we present a high-quality, large, and multilingual headline-generation dataset "Mukhyansh", comprising over 3.39 million news article-headline pairs across 8 Indian languages; namely Telugu, Tamil, Kannada, Malayalam, Hindi, Bengali, Marathi, and Gujarati. The data collection process for all eight Indian languages involved web scraping from multiple news websites as news websites and online news platforms serve as invaluable resources, offering a wealth of information that is instrumental in creating datasets for various NLP tasks. These platforms provide a rich and diverse pool of content, reflecting real-world behaviors and trends in news reporting. 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.

3.2.1 Preprocessing

Since the data is collected from various news websites, there is a possibility of same article-headline pairs present in the data collected from various websites, which leads to the problem of data duplication. It is important to eliminate duplicate article-headline pairs to ensure the integrity and quality of the dataset. This pre-processing step helps prevent biases in the model caused by redundant data and ensures that each article-headline pair contributes uniquely to the model's training process. We carry out a series of pre-processing steps including data de-duplication i.e. removing the duplicate article-headline pairs, and also ensure that there is no data overlap (contamination) across training, development, and test datasets.

We often encounter news articles whose headline is directly present in the article's prefix, this can be due to the style of writing headlines or an error due to automatic web crawling. When headlines are directly present in the article's prefix, models may learn to mimic this pattern rather than comprehensively understanding the article's content. This issue stems from training data biases, where models may prioritize copying existing headlines rather than generating contextually relevant summaries. To mitigate this, article-headline pairs with headlines directly present in the prefix of the article are removed from the dataset.

Another important issue present in the IndicHG dataset is that some articles contain multiple news article-headline pairs within the same piece of text. This phenomenon can be attributed to issues in data crawling, where other related or unrelated articles on a webpage become attached to a particular article. We eliminate these samples from our dataset as part of data preprocessing. Additionally, we have eliminated article-headline pairs where the article contains fewer than 20 words and/or the headline consists of fewer than 3 words. Table 3.2 provides an overview of the preprocessing statistics for Mukhyansh and the final Train, Dev, and Test splits.

3.2.2 Models & Results

The NLP community has proposed various modeling techniques for text generation tasks such as headline generation, spanning from traditional RNN-based approaches to more advanced transformer-based methods. We evaluate the performance of commonly used sequence-to-sequence models as baselines on the Mukhyansh dataset. Our implementation includes two categories of models: one based on an RNN encoder-decoder network trained from scratch, and another utilizing fine-tuning with pre-

	Dravidian language family				Indo-Aryan language family			
	te	ta	kn	ml	hi	bn	mr	gu
# Pairs collected	1080665	378545	505641	435896	729950	309008	411566	338502
# Duplicates	8024	11546	64116	269	32539	7055	10184	35518
# Pairs after deduplication	**1072641**	**366999**	**441525**	**435627**	**697411**	**301953**	**401382**	**302984**
# Pairs with prefix	8756	1712	1983	21633	2656	1302	942	200
# Pairs with multiple-articles	582	0	0	0	0	0	0	0
# Pairs too short	146181	33579	101619	98921	94132	19378	65998	26826
# Pairs after filtering	**917122**	**331708**	**337923**	**315072**	**600623**	**281273**	**334442**	**275958**
# Pairs in train	825372	298543	304122	283555	540568	253139	301001	248367
# Pairs in dev	82571	26539	27044	25190	48042	22514	26751	22073
# Pairs in test	9179	6626	6757	6327	12013	5620	6690	5518

Table 3.2 Statistics of Mukhyansh Dataset

trained transformer encoder-decoder models like mT5 [46] and IndicBART [7].

For the RNN architecture, we adopt the recurrent neural network proposed by [42], with a simple context attention mechanism inspired by [27], which is a modification of the dot product attention mechanism introduced by [28]. We explore two variations of this model: one using GRU [4] in both the encoder and decoder, and the other utilizing LSTM [20]. To tackle the challenge of out-of-vocabulary (OOV) words, particularly prevalent in morphologically rich Indian languages, we employ Byte Pair Encoding (BPE) [12]. Specifically, we use the GRU architecture[2] and initialize the model with 300d subword embeddings from BPEmb [19].

In addition to the approaches mentioned above, we also utilize transfer learning. This involves employing pre-trained sequence-to-sequence models such as mT5 [46] and IndicBART [7]. Transfer learning allows us to leverage knowledge acquired from extensive data and tasks, enabling these models to adapt effectively to the headline generation task with fine-tuning. By incorporating pre-trained models, we aim to enhance the performance and efficiency of our headline generation system. To implement these models, we utilize the scripts[3] provided by Huggingface [44].

Experimental Setup and Hyperparameters: The RNN-based models consist of 4 stacked layers, with each LSTM/GRU cell containing 600 hidden activation units. During the inference phase, we utilize the beam search strategy [45]. To prevent overfitting, we employ early stopping. Due to limited computational resources, for the pre-trained models we fine-tuned them on our data for 10 epochs. The model checkpoint with the highest validation score is selected to generate predictions on the test set. To assess the models' performance, we utilize the multilingual ROUGE metric [17][4]. Further details

[2]GRUs use fewer parameters, making them more computationally efficient for our experiments, with limited compute resources.

[3]https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization

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

regarding the experimental setup and parameter configurations for all the models can be found in Table 3.3.

Parameters	Seq-Seq + FastText	Seq-Seq + BPEmb	mT5-small	IndicBART
Max Source Length	200	300	1024	1024
Max Target Length	20	30	30	30
Vocabulary Size	40000	40000	250112	64000
Beam Width	5	5	4	4
Batch Size	16	16	16	16
Optimizer	Adam	Adam	Adam	Adam
Learning rate	$1e^{-4}$	$1e^{-4}$	$5e^{-5}$	$5e^{-5}$
(GPU,CPU)	(1,10)	(1,10)	(4,40)	(4,40)

Table 3.3 Experimental setup of various headline generation baseline models.

L	FastText+GRU			FastText+LSTM			BPEmb+GRU			mT5-small			IndicBART		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
te	32.71	15.00	32.02	33.41	14.93	32.70	30.06	14.52	29.31	39.34	21.95	**38.35**	38.42	20.85	37.33
ta	33.52	15.40	32.20	32.64	13.60	31.26	33.28	16.15	32.04	43.22	24.38	**41.18**	43.47	24.50	41.16
kn	26.19	10.53	25.25	23.75	7.94	22.84	24.46	10.68	23.60	34.73	17.88	**33.34**	34.36	17.06	32.59
ml	28.86	13.17	28.17	24.00	8.80	23.44	26.13	13.22	25.36	35.50	20.79	**34.63**	33.21	18.57	32.04
hi	32.97	14.20	29.50	32.34	11.79	28.45	32.24	13.93	28.94	38.26	18.81	33.65	41.05	20.77	**36.18**
bn	18.55	6.15	17.47	15.73	4.00	14.90	10.20	2.31	9.84	22.90	8.87	21.56	23.67	8.84	**22.04**
mr	17.26	5.08	16.83	14.32	3.11	14.04	17.91	6.48	17.54	27.25	12.68	26.41	28.21	12.95	**27.08**
gu	15.61	3.87	14.84	9.98	1.68	9.48	15.68	4.59	14.94	21.80	8.53	20.43	24.77	9.86	**23.05**
Average	25.71	10.43	24.54	23.27	8.23	22.14	23.75	10.24	22.70	32.88	16.74	31.19	33.40	16.68	31.43

Table 3.4 ROUGE-1,2,L scores of various headline generation baseline models trained on Mukhyansh dataset

Results: Table 3.4 presents the ROUGE-1, ROUGE-2, and ROUGE-L scores achieved by different baseline models on the Mukhyansh dataset. The best ROUGE-L score for each language is highlighted in bold. Notably, IndicBART and mT5-small emerged as the top-performing models, surpassing all sequence-to-sequence models trained from scratch. IndicBART achieved an average ROUGE-L score of 31.43 across all eight languages. This superior performance is attributed to their pre-training on a large corpus. The performance of IndicBART and mT5-small underscores the efficacy of transfer learning in headline generation task. By leveraging knowledge acquired from extensive data and tasks during pre-training, these models demonstrate a robust ability to generate headlines that closely resemble reference headlines. Furthermore, the GRU variant of the sequence-to-sequence model, using FastText word embeddings, produced decent results despite its smaller parameter count (64 million) compared to IndicBART (244 million) and mT5-small (300 million).

In summary, the results underscore the importance of pre-training and architectural choices in optimizing headline generation models. IndicBART and mT5-small demonstrate superior performance, emphasizing the benefits of transfer learning in capturing language nuances. Additionally, the competitive performance of the GRU variant highlights the significance of leveraging efficient embedding techniques and model architectures to achieve satisfactory results with reduced parameter counts.

3.3 TeClass Dataset

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. This section discusses the creation of the human-annotated dataset "TeClass". This dataset serves as a valuable resource for developing machine-learning models for classifying headlines based on their relevance to their corresponding news articles. Moreover, this dataset can be instrumental in examining how various categories of news headlines affect the performance of headline generation models. Previous studies on headline generation have often relied on training data obtained through web scraping, resulting in datasets comprising headlines from diverse categories with varying degrees of relevance. Hence, it's crucial to explore how the inclusion of different headline types impacts the performance of headline generation models. "TeClass" serves as a valuable resource for conducting such investigations.

3.3.1 Selecting the Article-Headline Pairs for Annotation

When selecting article-headline pairs for annotation, it's crucial to prioritize diversity in news content and coverage across various domains like sports, national news, entertainment, business, and more. This ensures that the resulting dataset represents a wide spectrum of topics and contexts, enabling the classification model to generalize effectively. Furthermore, to create a comprehensive and representative dataset, it's essential to include articles from a diverse set of news websites, encompassing both mainstream media outlets and niche publications, as well as regional and international sources. This approach accounts for variations in writing styles, editorial biases, and target audiences, enhancing the model's adaptability to different sources of news content. Keeping these factors into consideration we select a diverse subset of Telugu-article-headline pairs from "Mukhyansh" dataset which covers a vast set of news-domains and websites. In addition, we collect some article-headline pairs from the following two websites as well: Filmibeat[5], and Gulte[6] using web scraping.

3.3.2 Annotation Process

We employed crowd-sourcing for the annotation process, engaging native Telugu-speaking volunteers. The annotators were asked to assign one of the three primary categories: Highly Related (HREL), Moderately Related (MREL), and Least Related (LREL) after reading the headline and its corresponding article. They are also instructed to assign a secondary sub-class for each article-headline pair. The

[5] https://telugu.filmibeat.com/
[6] https://telugu.gulte.com/

following set of guidelines and instructions to the annotators are presented along with examples for assigning pre-defined categories to each article-headline pair.

Highly Related (HREL): The headline is highly related to the article content if it satisfies the following condition:

- 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
- *An example of highly related headline is presented in Figure 3.1*

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

- 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.
- *An example of moderately related headline is presented in Figure 3.2*

Least Related (LREL): The headline is least related to the article content if it satisfies any of the following conditions:

- 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.
- *An example of least related headline is presented in Figure 3.3*

Detailed Telugu examples for various fine-grained (secondary) categories of news headlines, along with explanations and translations are provided in subsequent pages.

3.3.2.1 Pilot Study

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.

3.3.2.2 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 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. We use Fleiss' Kappa metric proposed by [36] and it resulted in an encouragingly high score of 0.77, indicating a substantial agreement among the annotators.

3.3.3 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 1.1. 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 3.10, and Figure 3.11 respectively.

3.3.3.1 Data Splits

We allocated 70% of the data for training the model, 15% for development, and the remaining 15% data for testing purposes. To ensure unbiased performance and prevent category bias, we applied stratified sampling based on the category. This ensures even distribution of article-headline pairs from all 3 categories (HREL, MREL, LREL) across the training, development, and test sets. The category-wise counts in each data split are presented in Table 3.5. Further statistical details of the TeClass dataset are available in Table 3.6.

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 Vanithas signature was forged. The ministers signature was forged by a TDP leader from Kadapa district. Minister Thaneti Vanithas 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.

Figure 3.1 Example of Highly Related Headline

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

Table 3.5 Category-wise counts in each data split

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 విమర్శించారు. 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".

Figure 3.2 Example of 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 స్క్రిప్ట్లే 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	Upon reading the headline "Ranas wife as heroine in Mahesh Babus movie," readers may assume that the wife of Rana (in real life) is acting with Mahesh Babu. However, upon reading the article, it becomes apparent that the actress involved is actually Samyuktha Menon, not Rana's real-life wife. This discrepancy between the initial assumption created by the headline and the actual content of the article can mislead readers. This type of misleading conclusion is categorized as a least related headline.

Figure 3.3 Example of Least Related Headline

Factual Main Event Category Examples

Article: భోపాల్ : మధ్యప్రదేశ్లోని సింగ్రౌలి జిల్లాలోని కెర్హర్ గ్రామంలో బోరు బావిలో పడిన చిన్నారిని బయటకు తీశారు . ఆదివారం ఉదయం రెండేళ్ల చిన్నారి 70 అడుగుల లోతు గల బోరు బావిలో పడిన విషయం తెలిసిందే . చిన్నారిని సహాయక సిబ్బంది సురక్షితంగా రక్షించారు . చిన్నారిని రక్షించేందుకు పొక్లెయిన్ సాయంతో బోరుబావికి సమాంతరంగా పెద్ద గొయ్యిని తవ్వి చిన్నారిని అధికారులు బయటకు తీశారు . సహాయక సిబ్బంది వేగంగా స్పందించి చిన్నారిని కాపాడటంతో గ్రామస్థులు అభినందనలు తెలిపారు . పాప బతికి బయటపడడంతో చిన్నారి తల్లిదండ్రుల సంతోషానికి అవధులు లేకుండాపోయాయి.

Translation: Bhopal: A minor girl was pulled out of a borewell in Kerhar village of Madhya Pradesh's Singrauli district. On Sunday morning, a two-year-old girl fell into a 70-feet-deep borewell. The child was rescued safely by the rescue staff. To rescue the child, the authorities dug a large pit parallel to the borewell with the help of a poclain and pulled out the child. The villagers congratulated the rescue staff for responding quickly and saving the child. The joy of the child's parents knew no bounds as the baby survived.

Headline: బోరు బావి నుంచి సురక్షితంగా బయటపడిన చిన్నారి

Translation: Child rescued safely from borewell

Category: Factual main event

Explanation: The article mainly discusses the rescue of a child who fell into a borewell. The headline also conveys the same information.

Article: దుబ్బాక : సిద్దిపేట జిల్లా దుబ్బాక మున్సిపల్ ఎన్నికల సందర్భంగా మొదటి రోజు నామినేషన్లు పురస్కరించుకొని , మున్సిపల్ కార్యాలయం వద్ద బందోబస్తును పోలీస్ మున్సిపల్ ఎన్నికల నోడల్ అధికారి సిద్దిపేట ఏసీపీ రామేశ్వర్ పర్యవేక్షణ చేసారు . ఈ సందర్భంగా వారు మాట్లాడుతూ .. మున్సిపల్ ఎన్నికల షెడ్యూలు రాష్ట్ర ఎన్నికల సంఘం , నిర్దేశించిన ప్రకారం నామినేషన్లు వెయ్యాలని సూచించారు . నామినేషన్ దాఖలు చేసేటప్పుడు అభ్యర్థితో యుక్తంగా నలుగురిని మాత్రమే లోపలికి అనుమతిస్తామని , పర్మిషన్ లేకుండా ఎలాంటి ర్యాలీలు తీయకూడదని తెలిపారు . వేరే వారు ఎవరు లోపలికి రావద్దని సూచించారు . పోలీసు వారి సలహాలు సూచనలు పాటించాలని , వివిధ రాజకీయ పార్టీ అభ్యర్థులకు సూచించారు . దుబ్బాక మున్సిపాలిటీలో 20 వార్డులు , 41 పోలింగ్ బూతులు , 15 పోలింగ్ కేంద్రాలు కలవు . ఈ రోజు వరకు గత ఎన్నికల్లో కేసులు నమోదైన , శాంతి భద్రతలకు విఘాతం కలిగించే వారిని మున్సిపల్ ఎన్నికల పరిధిలో ఉన్న వారిని 28 మందిని బైండోవర్ చేయడం జరిగిందని తెలిపారు . ప్రశాంతమైన వాతావరణంలో నామినేషన్ల గురించి పటిష్టమైన బందోబస్తు ఏర్పాటు చేసినట్లు తెలిపారు . ఈ కార్యక్రమంలో దుబ్బాక సిఐ హరికృష్ణ , ఎస్సై స్వామి , భూంపల్లి ఎస్సై మమ్మద్ సర్దార్ జమాల్ , సర్కిల్ పోలీస్ సిబ్బంది పాల్గొన్నారు .

Translation: Siddipet: Siddipet ACP Rameshwar, the nodal officer for the municipal elections, supervised the bandobast at the municipal office on the first day of nominations for the Dubbaka municipal elections in Siddipet district. Speaking on the occasion, they said.. He suggested that nominations should be filed as per the schedule of municipal elections as prescribed by the State Election Commission. At the time of filing the nomination, only four persons will be allowed to enter the premises as per the capacity of the candidate and no rallies should be held without permission, he said. Others advised no one to come in. The police have advised candidates of various political parties to follow their suggestions and instructions. Dubbaka municipality has 20 wards, 41 polling booths and 15 polling stations. Till date, cases have been registered in the last elections and 28 people have been arrested for disrupting law and order in the municipal elections. He said a tight vigil has been put in place about the nominations in a peaceful atmosphere. Dubbaka CI Harikrishna, SI Swamy, Bhoompalli SI Mohammed Sardar Jamal and circle police personnel participated in the programme.

Headline: మున్సిపల్ ఎన్నికల సందర్భంగా పటిష్టమైన బందోబస్తు : ఏసీపీ

Translation: Tight security ahead of municipal elections: ACP

Category: Factual main event

Explanation: The article discusses tight security measures taken in view of dubbaka municipal elections which are explained by ACP. The headline is also conveying the same information.

Strong Conclusion Category Examples

Article: న్యూఢిల్లీ , డిసెంబర్ 9 : థియేటర్లలో సినిమా ప్రారంభానికి ముందు జాతీయ గీతం ప్రదర్శించాల్సిందేనని సుప్రీం కోర్టు స్పష్టం చేసింది . అయితే విషయంలో దివ్యాంగులకు కోర్టు మినహాయింపుఇచ్చింది . జాతీయ గీతం ప్రదర్శించేటప్పుడు దివ్యాంగులు లేచి నిలబడాల్సిన అవసరం లేదని జస్టిస్ దీపక్ మిశ్రా , జస్టిస్ అమిత్వారాయ్తో కూడిన ధర్మాసనం స్పష్టం చేసింది . అయితే ఎవరైనా ప్రశ్నిస్తే మాత్రం దివ్యాంగులు సంజ్ఞల ద్వారా తెలియజేయాల్సి ఉంటుందని బెంచ్ పేర్కొంది . థియేటర్లలో జనగణమన గీతం ప్రదర్శనకు సంబంధించి కేంద్రం పదిరోజుల్లో మార్గదర్శకాలు విడుదల చేస్తుందని అటార్నీ జనరల్ ముకుల్ రొహ్తి కోర్టుకు తెలిపారు.

Translation: New Delhi, Dec 9: The Supreme Court has made it clear that the national anthem should be played in theaters before the start of a film. However, the court made an exception to the divyangs in the matter. A bench of Justices Dipak Misra and AmitWarai made it clear that divyangs need not stand up while the national anthem is being played. The bench, however, said that if anyone asks a question, the differently-abled persons will have to inform through gestures. Attorney General Mukul Rohatgi told the court that the Centre will issue guidelines within 10 days regarding the screening of jana gana mana song in theaters.

Headline: జాతీయ గీతం తప్పనిసరి ..

Translation: The national anthem is mandatory..

Category: Strong conclusion

Explanation: The article mainly discusses the Supreme Court's order that the national anthem be played in movie theaters. The headline says “National Anthem is mandatory”. Although the headline misses the fact that it pertains specifically to movie theaters, one can infer the headline from the article and it supports the main event that is discussed in the article. Therefore, the headline can be considered a strong conclusion.

Article: తెలుగు ఆడియన్స్ ఎంతగానో ఎదురుచూస్తున్న సినిమాల్లో పవన్ కళ్యాణ్ 25వ చిత్రం కూడా ఉంది . కళ్యాణ్ కు బాగా కలిసొచ్చిన దర్శకుడు త్రివిక్రమ్ ఈ చిత్రాన్ని డైరెక్ట్ చేస్తుండటంతో ఇంతటి క్రేజ్ ఏర్పడింది . అభిమానుల్లోని ఈ క్రేజ్ మూలంగానే చిత్ర ప్రీ రిలీజ్ బిజినెస్ కళ్ళు చెదిరే రీతిలో జరుగుతోంది . ఇప్పటికే చిత్ర నైజాం డిస్ట్రిబ్యూషన్ హక్కులను రూ . 29 కోట్లకు దిల్ రాజు దక్కించుకోగా వర్సెస్ హక్కులు కూడా అదే స్థాయిలో రూ . 21 కోట్లు పలికాయి . బ్లూ స్కై సంస్థ ఈ భారీ మొత్తాన్ని చెల్లించింది . ‘బాహుబలి - 2’ తర్వాత ఇంత మొత్తం పలికిన తెలుగు సినిమా ఇదే కావడం చెప్పుకోదగిన విశేషం . ఈ మొత్తం తిరిగి రావాలంటే సినిమా మినిమమ్ హిట్ అవ్వాలి . ఇక సూపర్ హిట్టైతే లాభాలు కూడా భారీగానే ఉంటాయి . అలాగే చిత్ర శాటిలైట్ రైట్స్ రూ . 32 కోట్లు పలికాయట . ఈ స్థాయిలో ప్రీ రిలీజ్ బిజినెస్ జరుపుతోంది గనుకే ఓపెనింగ్స్ గ్రాండ్ గా ఉండాలనే ఉద్దేశ్యంతో 2018 జనవరి 10న చిత్రాన్ని సంక్రాంతి సీజన్లో రిలీజ్ చేయాలని నిర్ణయించారు . ఈ చిత్రానికి తమిళ సంగీత దర్శకుడు అనిరుద్ మ్యూజిక్ అందిస్తున్నారు .

Translation: Pawan Kalyan's 25th film is one of the most awaited films of the Telugu audience. Director Trivikram, who is very close to Kalyan, is directing this film and this craze has been created. Due to this craze among the fans, the pre-release business of the film is going on in an eye-catching manner. The film's Nizam distribution rights have already been acquired for Rs. Dil Raju was bought for Rs 29 crore and the rights were also sold for rs 29 crore at the same level. 21 crores. The Blue Sky company paid this huge amount. This is the first Telugu film after 'Baahubali 2'. For this amount to come back, the film has to be a minimum hit. If it is a super hit, the profits will also be huge. Also, the satellite rights of the film are priced at Rs. 32 crores. Since the pre-release business is going on at this level, it has been decided to release the film on January 10, 2018 in the Sankranthi season with the intention of making the openings grand. Tamil music director Anirudh Ravichander is composing the music for the film.

Headline: కళ్ళు చెదిరే రీతిలో ప్రీ రిలీజ్ బిజినెస్

Translation: Pre-release business in an eye-catching manner

Category: Strong conclusion

Explanation: The article mainly discusses the massive pre-release business of Pawan Kalyan’s 25th film directed by Trivikram. While the headline, "Pre-release business in an eye-catching manner," omits details about the specific movie it refers to (Pawan Kalyan’s 25th film), readers can infer this from the article. Despite this omission, the headline supports the main event discussed in the article. Therefore, the headline can be considered a strong conclusion.

Factual Secondary Event Category Examples

Article: కరోనా నుంచి కోలుకున్న ఉపరాష్ట్రపతి వెంకయ్యనాయుడు సోషల్ మీడియాలో ఆసక్తికరమైన పోస్టు చేశారు . కరోనా నుంచి కోలుకోవడం ఎంతో ఆనందదాయకం అని పేర్కొన్నారు . సెప్టెంబరు 29న కరోనా పాజిటివ్ అని తేలిన తర్వాత హోమ్ క్వారంటైన్ లోకి వెళ్లానని , ప్రస్తుతం పూర్తిగా కోలుకున్నానని వెల్లడించారు . అయితే , తన అర్ధాంగి ఉషకు కరోనా సోకకపోవడం పట్ల తాను ఎంతో సంతోషిస్తున్నానని తెలిపారు . ఆమె ఎంతో గుండెనిబ్బరంతో ఉందని , ఆమె ఆరోగ్యానికి వచ్చిన ఢోకా ఏమీ లేదని వెంకయ్య వివరించారు . అంతేకాకుండా , తన కార్యాలయంలో పనిచేస్తూ కరోనా వైరస్ ప్రభావానికి గురైన మరో 13 మంది ఉద్యోగులు కూడా పూర్తిగా కోలుకోవడం పట్ల కూడా అంతే సంతోషిస్తున్నానని వెంకయ్యనాయుడు తెలిపారు . " నా ఆరోగ్యం పట్ల ఎంతో జాగ్రత్త తీసుకున్న వైద్యులు , ఇతర వైద్య సిబ్బందికి కృతజ్ఞతలు తెలుపుకుంటున్నాను . వైద్య సిబ్బందికి వేళకు సలహాలు అందిస్తూ నా ఆరోగ్యం కుదుటపడడంలో తోడ్పాటు అందించిన ఎయిమ్స్ నిపుణులకు ధన్యవాదాలు . నేను నిజంగా వారి సేవల పట్ల సంతృప్తి చెందాను . ఇక , నాకోసం అహర్నిశలు పనిచేసిన నా వ్యక్తిగత సిబ్బంది విక్రాంత్ , చైతన్యలకు అభినందనలు " అంటూ వెంకయ్య ఫేస్ బుక్ లో పోస్టు చేశారు .

Translation: Vice President Venkaiah Naidu, who has recovered from covid-19, posted an interesting post on social media. It is a matter of great joy to be able to recover from corona. "I went into home quarantine after testing positive for coronavirus on September 29 and have now fully recovered. However, he said that he is very happy that his ardhangi Usha has not been infected with coronavirus. Venkaiah explained that she was very heartbroken and there was no dhoka that came to her health. Naidu also said that he was equally happy that 13 more employees who were affected by coronavirus while working in his office have also fully recovered. "I would like to thank the doctors and other medical staff for taking great care of my health. I thank the aiims experts for providing timely advice to the medical staff and helping me recover my health. I was really satisfied with their services. Congratulations to my personal staff Vikranth and Chaitanya who worked tirelessly for me," Venkaiah naidu wrote on Facebook.

Headline: మా ఆవిడకు కరోనా సోకనందుకు ఎంతో సంతోషంగా ఉంది : ఉపరాష్ట్రపతి వెంకయ్యనాయుడు

Translation: I am very happy that my wife did not get infected with coronavirus: Vice President Venkaiah Naidu

Category: Factual secondary event

Explanation: The majority of the article discusses Mr. Venkaiah Naidu's recovery from COVID-19 and his expressions of gratitude towards the people and medical staff who cared for him. Additionally, he mentions his happiness for his wife, who remained unaffected by the virus. This secondary information is reflected in the headline, so it can be considered a factual secondary event.

Article: టీడీపీ అధినేత , ఏపీ సీఎం నారా చంద్రబాబునాయుడి మనవడు దేవాన్ష్ కేశఖండన కార్యక్రమం పూర్తి అయ్యింది . నేటి ఉదయం తాత చంద్రబాబు చంకలో తల నిండా జుట్టుతో కనిపించిన ఆ బాలుడు కేశఖండన కార్యక్రమం తర్వాత తండ్రి నారా లోకేశ్ చేతిలో గుండుతో ప్రత్యక్షమయ్యాడు . దేవాన్ష్ పుట్టు వెంట్రుకల సమర్పణ కోసం చంద్రబాబు కుటుంబంతో సహా హిందూపురం ఎమ్మెల్యే నందమూరి బాలకృష్ణ కుటుంబం నిన్న సాయంత్రానికే చిత్తూరు జిల్లా చంద్రగిరి మండలం నారావారిపల్లెకు చేరుకుంది . నేటి ఉదయం ఇరు కుటుంబాలు బంధుమిత్రులతో కలిసి చంద్రబాబు కులదైవం నాగాలమ్మ గుడికి వెళ్లి ప్రత్యేక పూజలు చేశాయి . ఈ సందర్భంగా కొత్త బట్టలతో ముస్తాబైన దేవాన్ష్ తన తాత చంద్రబాబు చంకలో ముద్దులొలుకుతూ కనిపించాడు . దేవాన్ష్ ను ఎత్తుకుని ప్రత్యేకంగా తయారు చేయించిన గొడుగు కింద చంద్రబాబు కాలి నడకన నాగాలమ్మ గుడి చేరుకున్నారు . అనంతరం అరగంట పాటు సాగిన కేశఖండన కార్యక్రమం తర్వాత దేవాన్ష్ తన తండ్రి నారా లోకేశ్ చేతిలో గుండుతో ప్రత్యక్షమయ్యాడు .

Translation: Telugu Desam Party (TDP) chief and Andhra Pradesh Chief Minister N Chandrababu Naidu's grandson Devansh's hair-knitana programme has been completed. The boy, who was seen with his head full of hair in his grandfather Chandrababu's armpit this morning, appeared with a bullet in his father Nara Lokesh's hand after the hair-cutting ceremony. The family of Hindupuram MLA Nandamuri Balakrishna, including Chandrababu's family, reached Naravaripalle in Chandragiri mandal of Chittoor district yesterday evening for the presentation of Devansh's birth hair. This morning, both the families along with relatives and friends went to the temple of Nagalamma, the deity of Chandrababu's caste, and performed special pujas. On this occasion, Devansh, dressed in new clothes, was seen kissing his grandfather Chandrababu in the armpit. Chandrababu reached nagalamma temple on foot under a specially made umbrella carrying Devansh. After the half-an-hour-long hair-shaking ceremony, Devansh appeared with a bullet in the hands of his father Nara Lokesh.

Headline: నాగాలమ్మ గుడికి వెళ్లి ప్రత్యేక పూజలు చేసిన చంద్రబాబు

Translation: Chandrababu went to Nagalamma temple and performed special pujas

Category: Factual secondary event

Explanation: The majority of the article focuses on the haircut program (కేశఖండన కార్యక్రమం) of Chandrababu Naidu's grandson Devansh. Additionally, it touches upon some devotional events at a temple, as indicated in the headline. This can be considered a factual secondary event.

Weak Conclusion Category Examples

Article: ఇస్మార్ట్ శంకర్ సినిమా మంచి విజయం సాధించడంతో పూరి మళ్ళీ హిట్టు ట్రాక్ ఎక్కాడు . ఈ సినిమా బాక్స్ ఆఫీస్ వద్ద మంచి విజయాన్ని అందుకుంది . ఇప్పటికే ఈ సినిమా ఎనమిది రోజుల్లో 63 కోట్లు సాధించినట్లు నిర్మాత చార్మే తెలిపారు . సినిమా యూనిట్ కూడా ఆంధ్రప్రదేశ్ లోని పలు దియేటర్లో తిరుగుతూ ఫాన్స్ తో కలిసి సినిమా సక్సెస్ ని ఎంజాయ్ చేస్తున్నారు . అయితే తాజాగా సినిమాపైన ఓ ఛానల్ కి ఇంటర్వ్యూ ఇచ్చారు పూరి జగన్నాధ్ .. ఈ ఇంటర్వ్యూలో భాగంగా నేను చాలా విషయాల్లో బాద పడ్డానని అన్నారు . దర్శకుడుగా సంతోషంగా ఉన్నప్పటికీ నిర్మాతగా మాత్రం చాలా కష్టపడ్డాను అని పూరి చెప్పుకొచ్చాడు . ఇది నిర్మాతలకు సహజమేనని చెప్పుకొచ్చాడు . ఇంకా సినిమా రిలీజ్ కి ముందు మరింత టెన్షన్ ఉంటుందని , అసలు సినిమా విడుదల అవుతుందా లేదా అనే భయం కూడా ఉంటుదని వెల్లడించాడు . సినిమా హిట్టు అయితే పర్లేదు కానీ సినిమా ఫ్లాప్ అయితే మాత్రం బాధ మామూలుగా ఉండదని పూరి చెప్పుకొచ్చాడు . తనని ఈ క్రమంలో చాలా మంది ఇబ్బంది పెట్టారని వారి విషయాలను బయటపెడతానని పూరి చెప్పాడు . తనని ఎవరెవరు ఎంతెంత ఇబ్బంది పెట్టారో అన్ని డిపార్ట్మెంట్ లనుండి వెల్లడిస్తానని పూరి తెలియజేసాడు.

Translation: Puri is back on the hit track with the success of iSmart Shankar's film. The film was a good success at the box office. According to producer Charmme, the film has already earned Rs 63 crore in eight days. The film unit is also roaming around in various theaters in Andhra Pradesh and enjoying the success of the film along with the fans. However, Puri Jagannadh recently gave an interview to a channel on the film. As part of this interview, I felt bad about a lot of things. Though I am happy as a director, I have worked very hard as a producer," puri said. This is natural for the producers. He also revealed that there will be more tension before the release of the film and there will be a fear of whether the original film will be released or not. Puri said that it doesn't matter if the film is a hit, but if the film is a flop, the pain will not be the same. Puri said that he was harassed by a lot of people in the process and he would reveal their things. Puri said he would reveal from all departments who had harassed him and how much he had suffered.

Headline: త్వరలోనే అందరి విషయాలు బయటపెడతా : పూరి

Translation: I will reveal everyone's details soon: Puri

Category: Weak conclusion

Explanation: The article primarily discusses the success of the movie Ismart Shankar at the box office and director Puri Jagannath's interview comments, where he talks about the challenges faced as a producer during the film's release. However, the headline only reflects a small portion of the article that discusses his statement, 'Will reveal everyone’s details soon: Puri'. Therefore, the headline can be considered a weak conclusion as it does not fully represent the main information presented in the article.

Article: హైదరాబాద్ సింగరేణి కాలనీలో ఆరేళ్ల చిన్నారిపై అత్యాచారం చేసి , హత్య చేసిన ఘటన ఇరు తెలుగు రాష్ట్రాల్లో సంచలనాన్ని రేకెత్తించింది . ఈ ఘటనపై ప్రతి ఒక్కరూ తీవ్ర ఆవేదన వ్యక్తం చేస్తున్నారు . పలువురు నేతలు బాధిత కుటుంబాన్ని పరామర్శించారు . వైయస్సార్టీపీ అధ్యక్షురాలు షర్మిల కూడా బాధిత కుటుంబాన్ని పరామర్శించారు . అనంతరం ఆమె మాట్లాడుతూ , బాధిత కుటుంబానికి రూ . 10 కోట్ల పరిహారం చెల్లించాలని , నిందితుడిని కఠినంగా శిక్షించాలని రాష్ట్ర ప్రభుత్వాన్ని డిమాండ్ చేశారు . కేసీఆర్ ఇంట్లో కుక్క చనిపోతే ఒక అధికారిపై చర్య తీసుకున్నారని ... ఇప్పుడు ఒక చిన్నారి దారుణ హత్యకు గురైతే ముఖ్యమంత్రిలో చలనమే లేదని మండిపడ్డారు . బాధిత కుటుంబానికి న్యాయం జరిగేంత వరకు దీక్షకు దిగుతున్నానని ప్రకటించి , ఆ వెంటనే అక్కడే దీక్షకు కూర్చున్నారు . మరోవైపు షర్మిల తల్లి విజయమ్మ కూడా అక్కడకు చేరుకున్నారు . కూతురితో పాటు దీక్షలో కూర్చున్నారు .

Translation: Hyderabad: A six-year-old girl was allegedly raped and murdered in Singareni Colony in Hyderabad. Everyone is deeply saddened by the incident. Several leaders visited the victim's family. YSRCP president Sharmila also visited the victim's family. Later, she said, "The victim's family has been given Rs. He demanded that the state government pay a compensation of Rs 10 crore and punish the accused severely. If a dog dies in KCR's house, action will be taken against an officer. Now if a child is brutally murdered, there is no movement in the chief minister. He announced that he was going on a hunger strike until justice was done to the victim's family and immediately sat there for the fast. Sharmila's mother Vijayamma also reached the spot. He sat on a fast along with his daughter.

Headline: షర్మిలతో పాటు దీక్షలో కూర్చున్న విజయమ్మ

Translation: Vijayamma sat on the fast along with Sharmila.

Category: Weak conclusion

Explanation: The article mainly discusses the incident of rape and murder of a six-year-old girl in Singareni Colony of Hyderabad and YS Sharmila demanding the state government to provide compensation to the victim’s family and punish the accused. However, the headline only reflects a small portion of the article that mentions "Vijayamma sat in the deeksha along with Sharmila." Therefore, the headline can be considered a weak conclusion as it does not fully represent the main information presented in the article.

Misleading Conclusion Category Examples

Article: దేశం 75వ స్వాతంత్య్ర దినోత్సవ వేడుకల్ని ఘనంగా జరుపుకుంటోంది. దేశ రాజధాని ఢిల్లీలో ఎర్రకోట వేదికపై ప్రధాని నరేంద్ర మోదీ జాతీయ పతాకాన్ని ఆవిష్కరించారు. స్వాతంత్యం కోసం ప్రాణత్యాగం చేసినవారిని దేశం స్మరించుకుంటోందన్నారు. దేశ వ్యాప్తంగా 75వ స్వాతంత్య్ర దినోత్సవ వేడుకలు అత్యంత ఘనంగా, కోవిడ్ నిబంధనల మధ్య ప్రారంభమయ్యాయి. దేశ రాజధాని ఢిల్లీలో శత్రుదుర్భేద్యంగా మారిన ఎర్రకోట వేదికపై ప్రధానమంత్రి నరేంద్ర మోదీ జాతీయ పతాకాన్ని ఆవిష్కరించారు. ముందుగా రాజ్‌ఘాట్‌లోని మహాత్మాగాంధీ సమాధి వద్ద నివాళులర్పించి..అనంతరం జాతీయ పతాకాన్ని ఎగురవేశారు. ప్రధాని హోదాలో మోదీ 8వసారి జెండా ఎగురవేశారు. అనంతరం త్రివిధ దళాల నుంచి సైనిక వందనం స్వీకరించారు. వైమానిక దళ హెలికాప్టర్లు ఆకాశం నుంచి పూలవర్షం కురిపించారు. జెండా ఆవిష్కరణ అనంతరం దేశ ప్రజల్ని ఉద్దేశించి ప్రసంగించారు. ప్రజలందరికీ స్వాతంత్య్ర దినోత్సవ శుభాకాంక్షలు అందించారు. స్వాతంత్యం కోసం పోరాటం చేసిన త్యాగధనుల్ని దేశం స్మరించుకుంటోందన్నారు. దేశ సరిహద్దుల్లో పగలూ రాత్రి తేడా లేకుండా పహారా కాస్తున్న వీరజవాన్లకు ప్రణామాలు అర్పించారు ప్రధాని మోదీ. కరోనా మహమ్మారిపై వైద్యులు, సిబ్బంది చేసిన సేవలు ఎనలేనివని కీర్తించారు. ఒలింపిక్స్ లో భారత అథ్లెట్లు సత్తా చాటారాని..పతకాలు సాధించినవారంతా స్ఫూర్తి అని మోదీ చెప్పారు.

Translation: The country is celebrating its 75th Independence Day. Prime Minister Narendra Modi unfurled the national flag at the Red Fort in the national capital. The nation remembers those who sacrificed their lives for freedom. The 75th Independence Day celebrations across the country began with great pomp and gaiety amid covid-19 restrictions. New Delhi: Prime Minister Narendra Modi unfurled the national flag at the Red Fort in the national capital. Earlier in the day, he paid tributes at Mahatma Gandhi's samadhi at Rajghat. Later, the national flag was hoisted. Modi hoisted the national flag for the 8th time as prime minister. He then received a military salute from the three services. Air Force helicopters showered flowers from the sky. After the flag hoisting ceremony, he addressed the nation. He wished all the people a happy Independence Day. The country remembers the sacrifices who fought for freedom. Prime Minister Narendra Modi on Tuesday paid tributes to the brave soldiers guarding the country's borders day and night. He praised the services rendered by the doctors and staff on the coronavirus pandemic. Indian athletes have performed well at the Olympics. Modi said that all those who have won medals are an inspiration.

Headline: కరోనా కారణంగా నిరాడంబరంగానే స్వాతంత్ర్య దినోత్సవ వేడుకలు

Translation: Independence Day celebrations in a simple manner due to coronavirus

Category: Misleading conclusion

Explanation:
Fact: The article discusses the "Independence Day" celebrations at Red Fort, Delhi, in which Prime Minister Modi participated. The article also mentions that the celebrations are being conducted in a grand manner.
Misleading Info: The headline states, "Independence Day celebrations in a simple manner due to coronavirus," which is misleading because it diverges from the central information presented in the article.

Article: ముంబై : అనిల్ కపూర్ వారసురాలిగా తెరంగ్రేటం చేసిన హీరోయిన్ సోనమ్ కపూర్ ఇండస్ట్రీకి వచ్చిన దాదాపు ఆరేళ్లవైనే అయింది . అయితే అమ్మడుకి మాత్రం ఇప్పటి వరకు సరైన గుర్తింపు రాలేదు . ఏ ఒక్క స్టార్ హీరోతో నటించే అవకాశం రాక పోవడంతో స్టార్ హీరోయిన్ హోదా కూడా సొంతం చేసుకోలేక పోతోంది . అయితే సోనమ్ కపూర్కు 2014 సంవత్సరం ఒక గ్రేట్ ఇయర్గా మారబోతోంది . ఇప్పటికే ఆమె ఈ సంవత్సరం షారుక్ ఖాన్ హీరోగా తెరకెక్కే ' Raes ' చిత్రంలో అవకాశం దక్కించుకుంది . తాజాగా బాలీవుడ్ స్టార్ హీరో సల్మాన్ ఖాన్ సినిమాలో నటించే అవకాశం కూడా సొంతం చేసుకుంది . ఈ చిత్రానికి సూరజ్ భరతజ్య దర్శకత్వం వహించనున్నారు . దాదాపు 15 ఏళ్ల గ్యాప్ తర్వాత సల్మాన్ ఖాన్ , సూరజ్ భరతజ్య కాంబినేషన్లో సినిమా రాబోతోంది . వీరి చివరి చిత్రం ' హమ్ సాత్ సాత్ హై ' . రంయినా చిత్రంలో సోనమ్ కపూర్ నటన నచ్చడంతో సూజర్ ఆమెను తీసుకున్నట్లు తెలుస్తోంది . ఈ చిత్రంలో హీరో పాత్రతో పాటు హీరోయిన్ పాత్రకు కూడా సరైన ప్రాధాన్యం ఉంటుందట . ఈ చిత్రానికి బడే భయ్యా అనే టైటిల్ పరిశీలిస్తున్నట్లు తెలుస్తోంది . దర్శకుడు సూరజ్ ఆర్ భరత్వాజకు ఇది కెరీర్లో 7వ చిత్రం . త్వరలో ఈచిత్రానికి సంబంధించిన పూర్తి వివరాలు తెలియనున్నాయి . సినిమా గురించి అఫీషియల్ అనౌన్స్మెంట్ వచ్చేదాకా ఎలాంటి విషయాలు మాట్లడకూడదని సోనమ్ నిర్ణయించుకున్నట్లు ఆమె సన్నిహితులు అంటున్నారు .

Translation: Mumbai: It's been more than six years since Sonam Kapoor made her debut as Anil Kapoor's successor. However, she has not received proper recognition so far. She is not able to get the status of a star heroine as she does not get a chance to act with any one star hero. But 2014 is going to be a great year for Sonam Kapoor. She has already bagged the opportunity to star in Shah Rukh Khan's 'Raees' this year. She has also bagged the opportunity to act in Bollywood star Salman Khan's film. The film will be directed by Sooraj Bharatajya. After a gap of almost 15 years, Salman Khan and Sooraj Bharatajya are coming together for a film. Their last film was 'Hum Saath Saath Hain'. Apparently, Sussaur has roped in Sonam Kapoor as she liked her performance in 'Raanjhanaa'. Apart from the role of the hero in the film, the role of the heroine will also be given due importance. It seems that the title bade bhaiyya is being considered for the film. This is the 7th film in director Sooraj R Bharathyaja's career. The full details of the film will be revealed soon. Sources close to Sonam say that she has decided not to say anything about the film until an official announcement about the film is made.

Headline: సల్మాన్ ఖాన్ను దక్కించుకున్న సోనమ్ కపూర్ !

Translation: Sonam Kapoor bags Salman Khan!

Category: Misleading conclusion

Explanation:
Fact: The article discusses Sonam Kapoor's opportunity to act alongside Shah Rukh Khan and Salman Khan in their upcoming films. **Misleading Info:** The headline states, "Sonam Kapoor bags Salman Khan," which is misleading because it diverges from the core information presented in the article.

Sensational, Clickbait Category Examples

Article: దిల్లీ : పార్లమెంటు రెండో విడత బడ్జెట్ సమావేశాలు సోమవారం ప్రారంభం కాగా ఉత్తరాఖండ్లో రాష్ట్రపతి పాలనపై ఉభయసభల్లో కాంగ్రెస్ సభ్యులు నిరసన వ్యక్తం చేశారు . కేంద్రం అప్రజాస్వామిక విధానాలకు పాల్పడుతోందని కాంగ్రెస్ ఆరోపించింది . లోక్సభలో కాంగ్రెస్ నేత మల్లికార్జున ఖర్గే నేతృత్వంలో ఆ పార్టీ సభ్యులు ఆందోళన ప్రారంభించారు . రాజ్యసభలో కాంగ్రెస్ సభ్యులు వెల్లోకి దూసుకుపోవడంతో సభను రెండుసార్లు వాయిదా వేశారు . రాజ్యసభ రెండోసారి సమావేశమైనపుడు కూడా పరిస్థితిలో మార్పు లేకపోవడంతో సభను మధ్యాహ్నం 2 గంటల వరకూ వాయిదా వేస్తున్నట్లు డిప్యూటీ చైర్మన్ హమీద్ అన్సారీ ప్రకటించారు .

Translation: New Delhi: As the second phase of the Budget session of Parliament began on Monday, Congress members protested in both houses of Parliament against president's rule in Uttarakhand. The Congress accused the Centre of indulging in undemocratic practices. The protest was launched by congress members led by Leader of The Congress in the Lok Sabha Mallikarjun Kharge. The Rajya Sabha was adjourned twice as Congress members rushed to the well of the House. Deputy Chairman Hamid Ansari announced that the Rajya Sabha will be adjourned till 2 pm as the situation remained unchanged even when it met for the second time.

Headline: ఉత్తరాఖండ్పై దద్దరిల్లిన పార్లమెంటు

Translation: Parliament is in a state of chaos over Uttarakhand

Category: Sensational

Explanation: The article talks about congress party protesting in parliament against centre's decision to impose president's rule in Uttarakhand. The headline says "Parliament is in a state of chaos over Uttarakhand" , which uses strong language 'దద్దరిల్లిన పార్లమెంటు' to catch the attention of the reader and thereby making it sensational.

Article: యంగ్ టైగర్ ఎన్టీఆర్ నటిస్తున్న లేటెస్ట్ మూవీ 'ఆర్ఆర్ఆర్' ఇప్పటికే రిలీజ్కు రెడీ అయిన సంగతి తెలిసిందే . ఈ సినిమాతో మరోసారి బాక్సాఫీస్ వద్ద రికార్డుల సునామీని క్రియేట్ చేసేందుకు చిత్ర వర్గాలతో పాటు ప్రేక్షకులు కూడా ధీమా వ్యక్తం చేస్తున్నారు . ఇక ఈ సినిమాను దర్శకధీరుడు రాజమౌళి తెరకెక్కిస్తుండటంతో ఈ సినిమా బాక్సాఫీస్ వద్ద ఎలాంటి రికార్డులు క్రియేట్ చేస్తుందా అని అందరూ ఆసక్తిగా చూస్తున్నారు . అయితే ఈ సినిమా తరువాత తారక్ తన నెక్ట్స్ మూవీని మాటల మాంత్రికుడు త్రివిక్రమ్ శ్రీనివాస్ డైరెక్షన్లో తెరకెక్కించేందుకు రెడీ అవుతున్నాడు . కాగా ఈ సినిమాలో తారక్ లుక్ సరికొత్తగా ఉండబోతున్నట్లు తెలుస్తోంది . ఈ సినిమా తరువాత తారక్ తన నెక్ట్స్ మూవీని 'ఉప్పెన' చిత్ర దర్శకుడు బుచ్చిబాబు సానా దర్శకత్వంలో తెరకెక్కించేందుకు రెడీ అవుతున్నట్లు తెలుస్తోంది . ఇటీవల బుచ్చిబాబు సానా తారక్కు ఓ స్పోర్ట్స్ నేపథ్యంలో సాగే కథను వినిపించాడట . ఈ సినిమాలో తారక్ 60 ఏళ్ల వయస్కుడిగా ప్రేక్షకులకు కనిపించేందుకు రెడీ అవుతున్నాడట . ఇలా 60 ఏళ్ల వృద్ధుడిగా తారక్ కనిపించనుండటంతో ఈ సినిమాలో తారక్ ఎలా ఉండబోతున్నాడా అనే ఆసక్తి అప్పుడే సినీ వర్గాల్లో నెలకొంది . కాగా త్రివిక్రమ్ డైరెక్షన్లో తారక్ చేయబోయే సినిమాకు 'అయినను పోయిరావలె హస్తినకు' అనే టైటల్ను చిత్ర యూనిట్ కన్ఫం చేసినట్లు ఇండస్ట్రీ వర్గాల్లో టాక్ వినిపిస్తోంది . ఈ సినిమాకు సంబంధించిన షూటింగ్ను త్వరలో ప్రారంభించేందుకు చిత్ర యూనిట్ రెడీ అవుతోంది .

Translation: Young Tiger NTR's latest movie 'RRR' is already ready for release. The film fraternity as well as the audience are confident of creating a tsunami of records at the box office once again with this film. With Rajamouli directing the film, everyone is eagerly waiting to see if the film will create any records at the box office. After this film, Tarak is all set to direct his next movie under the direction of trivikram srinivas. It looks like Tarak's look in the film is going to be a brand new one. After this film, Tarak is all set to direct his next movie under the direction of 'Uppena' director Buchi Babu Sana. Recently, Buchi Babu narrated a story to Sana Tarak in the backdrop of a sport. Tarak is all set to appear as a 60-year-old in the film. Since Tarak will be seen as a 60-year-old man, there is a lot of interest in the film circles as to how Tarak is going to be in the film. Meanwhile, there is talk in industry circles that the film unit has confirmed the title of 'Ayanu Poiravale Hastinaku' for Tarak's upcoming film to be directed by Trivikram. The film unit is all set to start the shooting of the film soon.

Headline: 60 ఏళ్లకు చేరుతున్న తారక్ .. ఆందోళనలో ఫ్యాన్స్ !

Translation: Tarak turns 60 Fans are worried!

Category: Clickbait

Explanation: The article mainly discusses the upcoming movies of Jr. NTR and mentions that he is going to appear as a 60-year-old man in the film directed by Buchibabu. However, the headline, 'Tarak turns 60 years.. Fans are worried!', creates curiosity in the reader by providing misleading/sensational information that is not actually present in the article. This creates a disconnect between what is presented in the headline and what is actually present in the article, making it a clickbait headline.

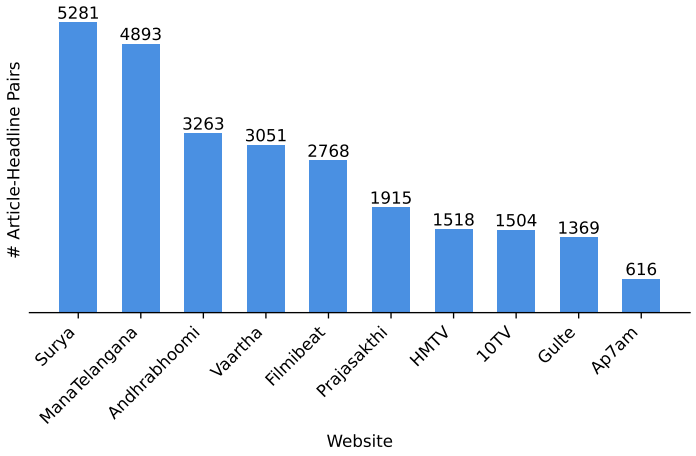

Figure 3.10 News website distribution in TeClass

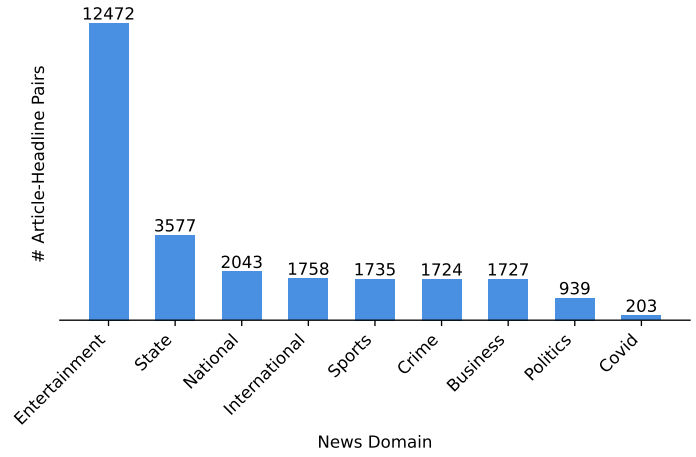

Figure 3.11 News domain distribution in TeClass

	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 words in article	126.33	126.70	126.39
Average words in headline	6.16	6.15	6.11
Unique words in articles	204959	76279	76070
Unique words 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 3.6 TeClass Statistics

Chapter 4

Relevance-based Headline Classification Experiments

In this Chapter, we discuss the implementation and evaluation of various baseline models for the task of classifying news headlines based on their relevance to their news articles. These baselines encompass traditional feature-based Machine Learning (ML) models for classification, as well as cutting-edge transfer learning techniques utilizing state-of-the-art pre-trained BERT [8] models.

4.1 Feature-based ML baseline models

We initiate our experimentation by using a suite of traditional feature-based Machine Learning (ML) models, leveraging a range of classification techniques including Logistic Regression (LR), Support Vector Machines (SVM), Multi-layer Perceptron (MLP), and Bagging. Various existing works on headline classification make use of features like n-gram overlap, cosine similarity between vector representations of the article, and the headline, and other hand-crafted features [15].

We also experiment with various features, and our model architecture presented in Figure 4.1 is similar to the one proposed by [37]. 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 [16][1] score between the headline and the first sentence of the article.

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 4.1.

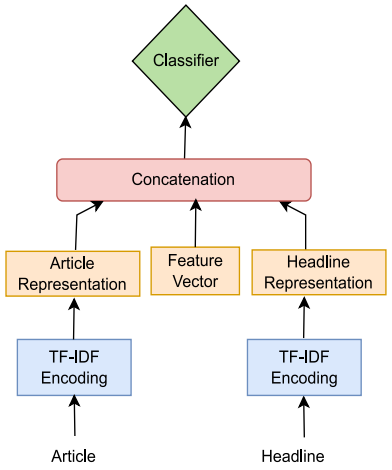

Figure 4.1 Feature-based machine learning baseline model architecture

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 4.1 Feature-based machine learning baseline model results

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

4.2 BERT-based baseline models

In this section we delve into the details of transfer learning, capitalizing on the transformative capabilities of pre-trained BERT models. BERT, short for Bidirectional Encoder Representations from Transformers, has emerged as a seminal advancement in natural language processing, offering contextual understanding and feature representation. 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. The model architecture is presented in Figure 4.2.

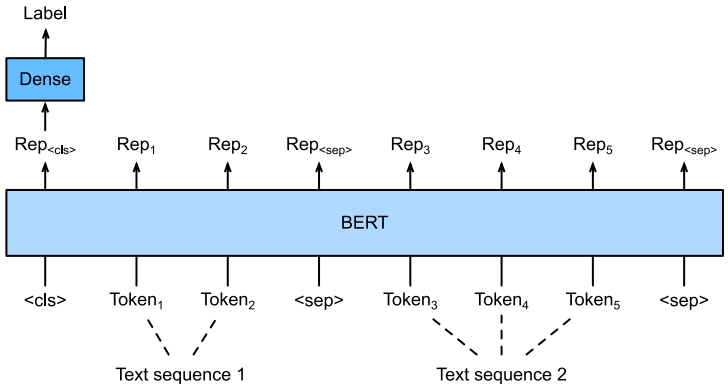

Figure 4.2 BERT-based baseline model architecture. Here, Text Sequence 1 is the Headline and Text Sequence 2 is its corresponding Article

We experiment with the following models by making use of the scripts [2] provided by Huggingface.

- **mBERT:** mBERT [8] is a multilingual variant of the BERT model. mBERT is a pre-trained model designed to support text processing in multiple languages (supports 102 different languages). It is trained using a masked language modeling (MLM) objective, where 15% of the words in a sentence are randomly masked, and the model predicts these masked words. Additionally, it employs a next sentence prediction (NSP) objective to learn relationships between pairs of sentences. This enables Multilingual BERT to generate bidirectional representations of text across a wide range of languages. For our baseline, we fine-tune the base version of mBERT having 110M parameters.

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

- **XLM-RoBERTa:** XLM-RoBERTa [6] is a state-of-the-art pretrained language model developed by Facebook AI. It is a multilingual version of the RoBERTa model, which is based on BERT architecture, but extends its capabilities to multilingual understanding. XLM-RoBERTa was pretrained on a vast 2.5TB CommonCrawl dataset, which included text from 100 languages, allowing it to effectively capture cross-lingual relationships and representations. For our experiments, we utilized the xlm-roberta-base variant, boasting 270 million parameters.

- **MuRIL:** The MuRIL [24] model utilizes a BERT base architecture trained from scratch on a diverse range of corpora, including Wikipedia, Common Crawl, PMINDIA, and Dakshina, encompassing 17 Indian languages. For training, it employs both monolingual and parallel data. Monolingual data comprises publicly available resources from Wikipedia and Common Crawl, while parallel data consists of translations and transliterations. Translations are obtained using the Google NMT pipeline and PMINDIA corpus, while transliterations are generated with the IndicTrans library from Wikipedia and Dakshina dataset. We employed the muril-base-cased variant with 236 million parameters for our task.

- **IndicBERT:** IndicBERT [9] 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 [18] is a multilingual adaptation of the DeBERTa model, maintaining the same architecture while trained on the CC100 multilingual dataset. The mDeBERTa V3 base model consists of 12 layers and a hidden size of 768. With 86 million backbone parameters, it incorporates a vocabulary of 250,000 tokens, resulting in 190 million parameters in the Embedding layer. This model was trained on 2.5 terabytes of CC100 data (featuring text from 100 languages), similar to the approach used for XLM-R. We used the base variant of mDeBERTaV3 in our experiments.

Hyperparameters: For all these BERT-based pre-trained 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 4.2.

4.3 Results & Analysis

ML Baselines: From the results presented in Table 4.1, 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 significant 3% improvement when compared to the model that did not utilize a feature vector.

BERT Baselines: The results presented in Table 4.2 underscore the superiority of state-of-the-art BERT-based models in comparison to classical machine learning models. The best model, mDeBERTa, achieved an 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.

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 4.2 BERT-based headline classification baseline model results

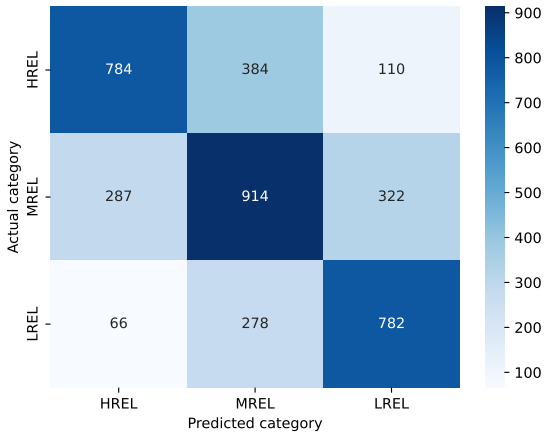

Figure 4.3 Confusion matrix between actual and predicted categories of mDeBERTa model

Challenges: The confusion matrix between actual categories and predicted categories of the mDeBERTa model shown in Figure 4.3 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 these classes. This observation sheds light on the complexities inherent in relevance-based headline classification tasks. It underscores the need for more sophisticated modeling techniques to address such challenges effectively.

Two-class classification: Since the models struggle to effectively differentiate between highly relevant and moderately relevant headlines, we can also approach this problem as a 2-class classification problem instead of a 3-class classification. For this purpose, we consider the fine-grained classes Factual Main Event (FME), Factual Secondary Event (FSE) and Strong Conclusion (STC) as Relevant to the article, and other classes including Weak Conclusion (WKC), Misleading Conclusion (MLC), Unsupported Personal Opinion (USO), Sensational (SEN) and Clickbait (CBT) as Less Relevant. We experiment with the same set of BERT-based pre-trained models for 2-class classification and the results are presented in Table 4.3. We can see 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.

Pre-trained model	**F1 Score**			
	FME+STC+FSE	WKC+MLC+SEN+USO+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 4.3 BERT-based headline classification baseline model results for merged fine classes

Chapter 5

Relevance-based Headline Generation Experiments

As discussed in earlier chapters, headline generation datasets collected from the web typically include a mixture of relevant and irrelevant headlines. However, there is a lack of studies focusing on how different types of headlines affect the performance of headline generation models. In this chapter, we present the details of several experiments conducted to investigate the impact of fine-tuning headline generation models using various categories of headlines, each with different degrees of relevance to the corresponding article. We experimented with a pre-trained mT5 headline generation model "Mukhyansh" [29], which was trained on a huge corpus of around 825k Telugu article-headline pairs for the task of headline generation. In our experiments, this model was further fine-tuned on different subsets of "TeClass" dataset to evaluate the impact of class-specific fine-tuning on the headline generation task.

5.1 Models & Results

- **Zero-shot inference**: We take the pre-trained headline generation model Mukhyansh and run zero-shot inference on news articles from various headline classes present in the test dataset of TeClass.
- **Fine-tune only on FME (Factual Main Event) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on FME class data present in the training dataset of TeClass.
- **Fine-tune only on STC (Strong Conclusion) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on STC class data present in the training dataset of TeClass.
- **Fine-tune only on FSE (Factual Secondary Event) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on FSE class data present in the training dataset of TeClass.

- **Fine-tune only on WKC (Weak Conclusion) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on WKC class data present in the training dataset of TeClass.
- **Fine-tune only on SEN (Sensational) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on SEN class data present in the training dataset of Teclass.
- **Fine-tune only on CBT (Clickbait) category data**: The pre-trained headline generation model Mukhyansh is fine-tuned only on CBT class data present in the training dataset of Teclass.
- **Fine-tune on all 6-class data**: The pre-trained headline generation model Mukhyansh is fine-tuned on (FME, STC, FSE, WKC, SEN, CBT) 6-class data present in the training dataset of TeClass.
- **Fine-tune on more relevant classes (FME, STC, FSE) data**: The pre-trained headline generation model Mukhyansh is fine-tuned on (FME, STC, FSE) 3-class data present in the training dataset of TeClass.
- **Fine-tune on less relevant classes (WKC, SEN, CBT) data**: The pre-trained headline generation model Mukhyansh is fine-tuned on (WKC, SEN, CBT) 3-class data present in the training dataset of TeClass.

The results of these models fine-tuned on various class-based data is presented in Table 5.1.

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

From the results in the Table 5.1, we can observe that the performance of the Zero-shot inference of the ”Mukhyansh” model (pre-trained on a huge Telugu corpus) falls short when compared to the model that is fine-tuned only on FME data. The model fine-tuned on FME data significantly outperformed the zero-shot inference model by 6 ROUGE-L points when tested on FME, FSE category data of TeClass, and 5 ROUGE-L points when tested on STC category data. In summary, fine-tuning only on FME category data always significantly improves ROUGE-L scores (5 points on average) across different categories.

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 three relevant classes (FME, STC, FSE). It is also interesting to see that the performance gain is not proportional to the training data size. We see a marked decrease in performance when all of the 6-class data is used for fine-tuning. The best performance is achieved by fine-tuning on Factual Main Event (FME) data that constitutes only 43% of the total training data of "TeClass" dataset. It shows that even though the highly related data is relatively small compared to the total dataset size, fine-tuning only on the highly related data significantly performs better than fine-tuning on the entire data (containing highly relevant, moderately relevant, and least relevant data). It is noteworthy to mention that, in addition to the performance gain, fine-tuning only on highly relevant data saves compute resources when compared to fine-tuning on the entire dataset.

In conclusion, the findings highlight the pivotal role of fine-tuning headline generation models on highly related headlines, significantly enhancing their performance. This underscores the importance of leveraging relevance-based headline classification datasets, like the "TeClass" dataset, to facilitate the generation of highly pertinent headlines.

Chapter 6

Conclusions

In this work, we present "Mukhyansh", a large, diverse multilingual headline generation dataset containing 3.39 Million news article-headline pairs across 8 Indian languages. We provide various baseline models for the task of headline generation and evaluate their performance. In addition, we introduce a novel, high-quality human-annotated dataset named "TeClass" 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 (HREL), Moderately Related (MREL), and Least Related (LREL). 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 relevance-based headline classification in the Telugu language.

In our experiments with various baseline models on the "TeClass" dataset for relevance-based headline classification, our empirical findings highlight the superior performance of BERT-based models when compared to classical machine learning models. 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.

Furthermore, we leverage the "TeClass" dataset to examine the impact of fine-tuning headline generation models on different classes of headlines with varying degrees of relevance. Our experiments demonstrate that fine-tuning solely on highly relevant data yields significantly better performance compared to fine-tuning on the entire dataset containing a mix of all categories of data. This approach also helps conserve computational resources required for model training.

Publications

Relevant Publications

1. **Gopichand Kanumolu**, Lokesh Madasu, Nirmal Surange, Manish Shrivastava. "TeClass: A Human-Annotated Relevance-based Headline Classification and Generation Dataset for Telugu." In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

2. Lokesh Madasu, **Gopichand Kanumolu**, Nirmal Surange, Manish Shrivastava. "Mukhyansh: A Headline Generation Dataset for Indic Languages." In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 2023).

Other Publications

1. **Gopichand Kanumolu**, Lokesh Madasu, Pavan Baswani, Ananya Mukherjee and Manish Shrivastava, "Unsupervised Approach to Evaluate Sentence-Level Fluency: Do We Really Need Reference?". In Proceedings of the First Workshop in South East Asian Language Processing, IJCNLP-AACL 2023.

2. Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine de Kock, Genet Shanko Dekebo, Oumaima Hourrane, **Gopichand Kanumolu**, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, Saif M. Mohammad, "SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 14 Languages". In Findings of the Association for Computational Linguistics: ACL 2024.

Bibliography

[1] Michele Banko, Vibhu O Mittal, and Michael J Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325, 2000.

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

[3] Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 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, September 2017. Association for Computational Linguistics.

[4] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[5] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98, 2016.

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

[7] Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*, 2021.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 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, 2023.

[10] Bonnie Dorr, David Zajic, and Richard Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. Technical report, MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2003.

[11] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *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, 2016.

[12] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

[13] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

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

[15] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 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, August 2018. Association for Computational Linguistics.

[16] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics.

[17] Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*, 2021.

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

[19] Benjamin Heinzerling and Michael Strube. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. *arXiv preprint arXiv:1710.02187*, 2017.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[21] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 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, 2020.

[22] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*, 2020.

[23] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.

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

[25] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Comput. Surv.*, 53(1), feb 2020.

[26] Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M Khapra, and Pratyush Kumar. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages. *arXiv preprint arXiv:2203.05437*, 2022.

[27] Konstantin Lopyrev. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712*, 2015.

[28] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[29] Lokesh Madasu, Gopichand Kanumolu, Nirmal Surange, and Manish Shrivastava. Mukhyansh: A headline generation dataset for indic languages. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 620–634, 2023.

[30] Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *Transactions on Asian and Low-Resource Language Information Processing*.

[31] Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. Clickbait detection in telugu: Overcoming nlp challenges in resource-poor languages using benchmarked techniques. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[32] Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *arXiv preprint arXiv:2401.02254*, 2024.

[33] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[34] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

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

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

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

[38] Alexander M. Rush, Sumit Chopra, and Jason Weston. 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, September 2015. Association for Computational Linguistics.

[39] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[40] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[41] V Srinidhi Skanda, M Anand Kumar, and KP Soman. Detecting stance in kannada social media code-mixed text using sentence embedding. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 964–969. IEEE, 2017.

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[43] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. An english-hindi code-mixed corpus: Stance annotation and baseline system. *arXiv preprint arXiv:1805.11868*, 2018.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[45] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[46] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.