Headline Generation for Indian Languages

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

Master of Science in
Computer Science and Engineering
by Research

by

LOKESH MADASU
2021701042
lokesh.madasu@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled *"Headline Generation for Indian Languages"* by Lokesh Madasu, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Manish Shrivastava

To my family and friends

Acknowledgements

I am deeply grateful to my advisor, Dr. Manish Shrivastava, for encouraging my passion for NLP and for his constant guidance and support throughout my journey as a master's student. His mentorship has been instrumental in shaping my success, and I am truly fortunate to have had the opportunity to work under his supervision. His encouragement and trust in my abilities have empowered me to explore various fascinating projects and have inspired me to realize my full potential.

Special thanks to my friends and collaborators, Gopichand Kanumolu and Pavan Baswani, for their invaluable contributions to my research journey at IIIT. Together, we have tackled challenging tasks and embarked on numerous research projects, producing impactful outcomes. I am deeply appreciative of their dedication and teamwork, which have created a supportive environment for our collaborative work.

I would like to extend a special thanks to my senior, Ashok Urlana, for generously dedicating his valuable time to resolving my various queries throughout my master's program. He serves as a perfect example of a supportive senior who consistently guides juniors like me along the right path. His research attitude always inspires me to continue in research.

I would also like to extend my gratitude to LTRC senior PhD students, Nirmal Surange, Ananya Mukherjee, Prashanth Kodali, Hiranmai Adibatla, and Aparajitha, for their valuable guidance and suggestions. They are always supportive and helpful.

Lastly, I am deeply grateful to my family members: my father, Mr. Ramakrishna Madasu; mother, Mrs. Siva Kumari Madasu; and brother, Jagadeesh Madasu, for their boundless love and constant support. Their continuous encouragement and the freedom they granted me motivated me to explore life in the best possible ways. Without their constant encouragement, I would not have reached this point.

Thank you IIIT Hyderabad for the invaluable experiences.

Abstract

In the field of Natural Language Processing (NLP), the abundance of online content presents both opportunities and challenges. The internet hosts a wealth of information, encompassing diverse topics and languages, ranging from news articles to blog posts. However, navigating through this sheer volume of content can be overwhelming, leading to information overload and difficulty in identifying the most relevant content. As a result, there is a growing demand for efficient methods to distill complex textual information into concise and informative summaries. One key approach to addressing this challenge is through headline generation.

Headline generation within the domain of NLP holds immense significance, particularly in today's era of short attention spans and overwhelming information flow. The ability to quickly grasp the key points of a document can significantly enhance user experience and facilitate knowledge dissemination, especially across diverse linguistic communities. Despite considerable advancements in headline generation for widely spoken languages like English, challenges persist in generating headlines for low-resource languages, such as the rich and diverse Indian languages. One major obstacle hindering headline generation in Indian languages is the limited availability of high-quality data.

To address this crucial gap, we introduce Mukhyansh, an extensive multilingual dataset tailored for Indian language headline generation. Mukhyansh comprises over 3.39 million article-headline pairs collected from the web, covering eight prominent Indian languages: Telugu, Tamil, Kannada, Malayalam, Hindi, Bengali, Marathi, and Gujarati. This thesis presents a comprehensive evaluation of several state-of-the-art baseline models on the Mukhyansh dataset. Through empirical analysis, we demonstrate that Mukhyansh surpasses existing models, achieving an impressive average ROUGE-L score of 31.43 across all eight languages.

However, the presence of irrelevant headlines in scraped news articles results in the sub-optimal performance of headline generation models. We propose that relevance-based headline classification can greatly aid the task of generating relevant headlines. Relevance-based headline classification involves categorizing news headlines based on their relevance to the corresponding news articles. While this task is well-established in English, it remains under-explored in low-resource languages like Telugu due to a lack of annotated data.

To address this gap, we present "TeClass", the first-ever human-annotated relevance-based news headline classification dataset for Telugu, containing 78,534 annotations across 26,178 article-headline pairs. We use this data set to demonstrate the impact of fine-tuning headline generation models on

various categories of headlines (with varying degrees of relevance to the article) and prove that the task of relevant headline generation is best served when the models are fine-tuned on a dataset containing highly relevant headlines, even though the size of highly related data is less in number.

Our work highlights the effectiveness of Mukhyansh and TeClass in advancing headline generation and classification research for Indian languages and underscores its potential to facilitate further developments in this domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's fast-paced digital world, staying informed is crucial, with an overwhelming amount of news and updates available at our fingertips. However, amidst this flood of information, finding the most relevant and important news can be like searching for a needle in a haystack. This is where headline generation steps in, offering a concise and informative summary of news articles that helps readers quickly understand what's happening in the world. Despite its potential benefits, headline generation comes with its own set of challenges. Headlines need to be informative yet engaging, striking a balance between providing essential information and sparking readers' interest. Also, headline generation systems must be capable of dealing with the differences in languages, meeting the needs of various language users. The objective of this thesis is to outline the motivation behind developing headline generation resources and systems, specifically tailored for Indian languages. Ultimately, our goal is to contribute to the advancement of natural language processing applications and ensure that everyone has access to relevant and engaging news content.

1.1 Motivation

Natural Language Generation (NLG) is a subfield of Natural Language Processing (NLP) concerned with the automatic generation of human-like text. Abstractive summarization is a specific task within NLG that involves generating a condensed and meaningful representation of a longer text. Notably, headline generation for news articles can be seen as a form of abstractive summarization, where the objective is to craft a single-sentence summary that accurately encapsulates the article's content.

The task of headline generation is particularly challenging, since the headline must be both relevant and creative. It plays a crucial role in summarizing news articles and capturing readers' attention. This task involves automatically generating informative and captivating headlines that accurately capture the essence of the underlying text. Headline generation is challenging due to two major factors. Firstly, headlines must accurately represent the content of the text while being concise. Secondly, headlines often need to be attention-grabbing, compelling readers to click and read further. This thesis presents

resources and systems for generating headlines in Indian languages and also focuses on generating relevant headlines specifically in Telugu.

1.2 Headline Generation in Indian Languages

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 realm of headline generation for Indian languages. While datasets like Gigaword [23] have emerged as prominent resources, comprising an impressive collection of over 4 million news article-headline pairs, it is crucial to acknowledge that they are limited to English and fail to capture the intricacies and linguistic nuances of Indian languages.

Numerous datasets exist for training and assessing headline generation and abstractive summarization models, primarily focusing on English [24, 22, 27, 33]. While efforts like MLSum [28] and XLSum [14] have aimed to develop multilingual datasets, representation of Indic languages remains minimal. This scarcity poses challenges in creating effective headline generation systems for Indian languages.

Indic headline generation presents intriguing challenges due to the unique characteristics of these languages. Despite their linguistic similarities, the diverse scripts used across the Indic language family pose obstacles to effective transfer learning. Additionally, the morphological complexity of many Indic languages results in compact headlines, making prediction more difficult. Moreover, the majority of Indic languages suffer from low-resource status and are often overlooked in multilingual models, limiting the effectiveness of few-shot learning approaches. Recent studies have highlighted the importance of language family-specific pretraining for enhancing transfer learning capabilities [7], emphasizing the need for ample high-quality data to achieve optimal performance in downstream tasks such as text generation and summarization [36].

One of the most significant obstacles hindering headline generation in Indian languages is the scarcity of high-quality annotated data. This scarcity severely limits the effectiveness of model training and impedes the performance of supervised learning approaches, which heavily rely on labeled examples.

Fortunately, recent advancements in neural network architectures, such as transformer-based models IndicBART [6], mBART [20] and mT5 [35] have significantly enhanced the performance of headline generation models. These models possess the ability to encode input text and generate headlines by optimizing various objectives, including semantic coherence, informativeness, and readability. While these models have successfully reduced the dependency on labeled data, they still leverage fine-tuning on specialized headline generation datasets to further enhance their performance.

1.3 Relevance-based Headline Generation for Telugu

In today's digital landscape, headline act as a filter, allowing the reader to quickly decide if the story is relevant or interesting to them. 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. However, in the quest for clicks, sensationalism often takes precedence over accuracy, leading to the widespread dissemination of misinformation and fake news. This not only erodes public trust in media but also has tangible consequences, shaping individuals' decisions across a range of issues from personal health to political beliefs.

To counteract this trend, researchers are actively developing machine learning models capable of detecting and classifying misleading information at scale. These efforts are vital for upholding the integrity of information dissemination and ensuring accuracy and transparency in journalism.

Generating relevant headlines for news articles is quite a daunting task, mainly because of the varied quality and types of data used for training. When the dataset includes a mix of relevant headlines along with clickbait, sensational, or misleading ones, it can influence the model to generate similar types of headlines. Thus, it's crucial to have high-quality and relevant training data to train headline generation models effectively. These models should be able to produce accurate and meaningful headlines that inform and engage readers without resorting to sensationalism or misleading tactics.

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 [27, 13, 3]. 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 [17].

We believe that the generation of relevant headlines is contingent on the quality of the data presented to the models during training. We have observed that for low-resource languages like Telugu, the ratio of highly relevant headlines versus not-so-relevant or irrelevant headlines is badly skewed towards irrelevance (Figure 1.1). 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 substantially enhance the relevance and quality of the generated headlines.

However, progress in relevance-based headline generation field is hampered by the scarcity of high-quality datasets, particularly for the Telugu language. To address this critical gap, we introduce "TeClass", a human-annotated dataset designed for relevance-based headline classification in Telugu. With careful annotations categorizing article-headline pairs into Highly Related (HREL), Moderately Related (MREL), and Least Related (LREL) categories, TeClass serves as a foundational resource for

[1]Headline need not be a complete sentence

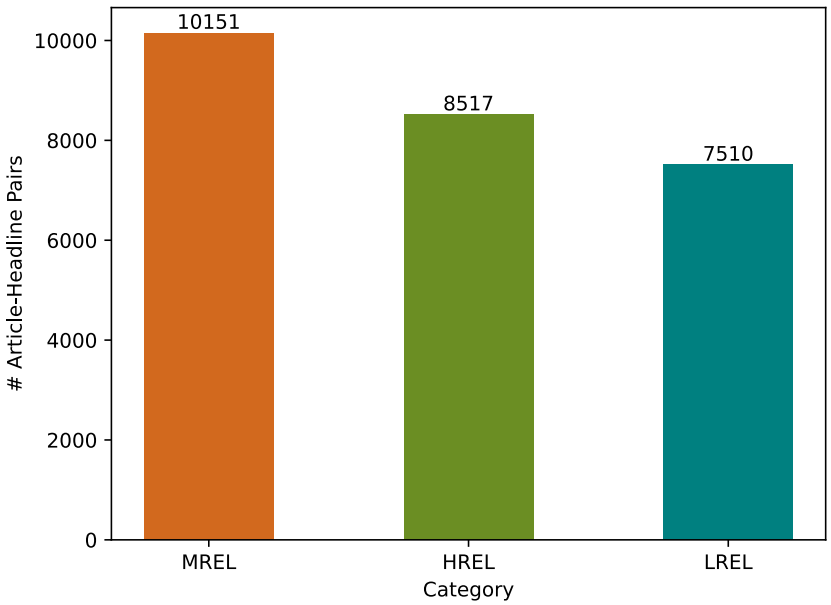

Figure 1.1: Category distribution in TeClass. HREL: High Relevance, MREL: Medium Relevance, LREL: Low Relevance

training and evaluating machine learning models tailored to the intricacies of the Telugu language and news media landscape.

1.4 Thesis Contribution

Following are the main contributions of this thesis.

1. We present a large, 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. Our data collection methodology involves developing site-specific crawlers, leveraging a deep understanding of news website structures to ensure the acquisition of high-quality data.

    (a) We employ state-of-the-art baseline models and demonstrate the effectiveness of these models for a diverse range of test sets.

    (b) We provide further evidence to support our argument regarding the necessity of high-quality data by undertaking a comprehensive comparative analysis, specifically contrasting our research with the existing work, particularly IndicHG.

(c) The dataset and models are available at: https://github.com/ltrc/Mukhyansh

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 headline classification with one of the three categories: Highly Related(HREL), Moderately Related(MREL) and Least Related(LREL).

   (a) We present baseline headline generation models to demonstrate that the task of relevant headline generation is best served when the generation models are trained on high-quality relevant data even if the available relevant article-headline pairs are significantly less in number.

   (b) The dataset and models are available at: https://github.com/ltrc/TeClass

1.5 Organization of thesis

- Chapter 1 provides an introduction, motivation for the headline generation task in Indian languages, and outlines the key contributions of this thesis.
- Chapter 2 discusses various modeling approaches and existing datasets for both English and multiple languages.
- Chapter 3 details the creation of Mukhyansh and evaluates the performance of various baseline models. Also, presents a comprehensive analysis of existing datasets, conducts a comparative assessment of each model's performance on diverse datasets, and presents our findings.
- Chapter 4 discusses the creation of TeClass and its importance in generating relevant headlines in low-resource language, Telugu.
- Finally, Chapter 5 summarizes the contributions and limitations of this thesis, and discusses its future scope.

Chapter 2

Literature Review

Generating headlines in low-resource languages has been a pivotal problem in the field of Natural Language Processing. Various methods have been proposed, from traditional statistical and linguistic approaches to advanced machine learning and deep learning techniques. In this chapter, we discuss various modeling approaches and existing datasets proposed for headline generation.

2.1 Modeling Approaches

The "parse-and-trim" approach [8] involves parsing the article into its constituent parts using a dependency parser and selecting the most relevant parts to form the headline using linguistically-motivated heuristics. Alternatively, a statistical approach proposed in [2] considers the problem of headline generation as similar to statistical machine translation. This approach utilizes a language model to predict the words in the headline given the words in the source article. The integration of neural networks and deep learning into NLG has brought about a significant shift in the field, resulting in more sophisticated and effective models capable of generating high-quality text. For instance, an attention mechanism-based summarization (ABS) method introduced in [27] employs a convolutional network in the encoder to encode the input words and utilizes attention to aid the text generation model. While this approach employs a feed-forward neural network for text generation, [5] proposed a Recurrent Neural Network (RNN) based decoder that demonstrated improved performance compared to the previous approach on the English Gigaword dataset. Additionally, Nallapati et al. [22] introduced a novel Encoder-Decoder network with a hierarchical attention mechanism employing recurrent neural networks in both the encoder and decoder. Gu et al. [12] proposed the COPYNET mechanism, enabling the model to generate text by copying words from the input, effectively addressing the out-of-vocabulary (OOV) words issue. Another approach, presented in [29], introduced a pointer generator network capable of generating a probability distribution over vocabulary and copying words from the input using attention distribution, thus effectively handling the OOV words problem. Furthermore, the authors introduced a novel coverage mechanism to mitigate the issue of repetitive words on the decoder side. In contrast to previous approaches aimed at generating plain and factual headlines, [17] developed a stylistic headline

generation system capable of generating headlines with three target styles: humor, romance, and clickbait. Recently, pretrained language models like BART [19] and T5 [25] have demonstrated significant success in headline generation and related tasks.

2.2 Existing Datasets

2.2.1 English

DUC 2003 and 2004 The DUC (Document Understanding Conference) datasets from 2003 and 2004 are widely used benchmark datasets in the field of text summarization. The DUC 2003 dataset comprises 500 article-headline pairs from news articles published by the Associated Press and the New York Times, covering various topics such as politics, sports, and entertainment. Similarly, the DUC 2004 dataset contains 624 article-headline pairs from the same sources, providing a larger and more diverse collection of pairs compared to the previous year. Both datasets are designed for single-document summarization tasks, with each article accompanied by a manually written headline that serves as a concise summary of the main points. These datasets have been instrumental in evaluating the effectiveness of summarization algorithms and approaches, allowing researchers to assess the performance of different systems in generating informative summaries from news articles.

Gigaword [27] The English Gigaword dataset comprises a vast collection of news articles, totaling over 4 million in number. These articles are sourced from reputable newspapers such as the New York Times and the Associated Press, covering a diverse range of topics spanning politics, economics, sports, entertainment, and more. Each article in the dataset is paired with its corresponding headline, providing labeled data for tasks such as headline generation and text summarization. Due to its extensive coverage and high-quality content, the English Gigaword dataset has become a fundamental resource for natural language processing research.

XSUM [24] The Extreme Summarization (XSum) dataset is designed to assess the performance of abstractive single-document summarization models. Comprising 226,711 news articles sourced from BBC articles spanning the years 2010 to 2017, the dataset covers a broad spectrum of domains including News, Politics, Sports, Weather, Business, Technology, Science, Health, Family, Education, Entertainment, and Arts. Each article is accompanied by a corresponding one-sentence summary, making XSum a valuable resource for evaluating and benchmarking abstractive summarization systems.

TLDR [33] The TLDR corpus is a dataset designed for summarizing Reddit posts. It consists of diverse user-generated content from the social media platform Reddit, covering a wide range of topics and discussions. The dataset contains both the original posts and corresponding human-generated summaries, providing valuable resources for training and evaluating summarization models. With its large-scale and varied content, the TLDR corpus serves as a valuable asset for researchers interested in developing and testing summarization algorithms on real-world social media data.

2.2.2 Multilingual

Columbia Newsblaster [9] The Columbia Newsblaster dataset is a multilingual dataset created for the purpose of summarization. It includes a collection of news articles in multiple languages, such as English, Russian, and Japanese, along with annotations for various elements including articles, texts, titles, images, and captions. The dataset encompasses around 500 news articles and spans a diverse range of topics and domains. This multilingual nature of the dataset makes it a valuable resource for researchers exploring summarization techniques across different languages and cultural contexts.

MLSum [28] It consists of approximately 1.5 million pairs of articles and summaries along with the headlines written in languages including French, German, Spanish, Russian, and Turkish. The dataset covers a wide array of topics, drawing from diverse domains such as news, politics, sports, weather, business, technology, science, health, family, education, entertainment, and arts.

XLSum [14] It comprises approximately 1 million article-summary pairs extracted from BBC news articles spanning various languages, including 44 different languages. The dataset aims to provide a diverse range of linguistic data for evaluating summarization systems across a wide linguistic spectrum. By including headlines along with their corresponding summaries, XLSum facilitates the assessment of headline generation models' performance in capturing the essence of news articles in different languages.

IndicNLG-HG To further advance research in Natural Language Generation (NLG) for Indian languages,[18] proposed the IndicNLG benchmark, encompassing five different NLG tasks, including a headline generation dataset (hereafter referred to as IndicHG dataset). This dataset comprises 1.31 million article-headline pairs across 11 Indian languages.

However, our analysis (detailed in Chapter 3 Section 3.3) reveals serious quality issues in IndicHG, such as data contamination, rendering it unsuitable for training robust models. Despite its claimed size, the dataset's problematic samples significantly reduce its effective size by nearly half.

Chapter 3

Mukhyansh

This chapter explores the creation of the headline generation dataset "Mukhyansh" spanning eight Indian languages: Telugu, Tamil, Kannada, Malayalam, Hindi, Bengali, Marathi, and Gujarati. We examine the performance of various baseline models using this dataset, along with a critical evaluation of existing datasets. Additionally, we also presents comparative analysis of each model's performance across different datasets.

3.1 Introduction

News websites provide a wealth of up-to-date and diverse content, with news stories covering a wide range of topics and from various perspectives. This variety can be valuable for training NLG models to generate natural and diverse language which is representative of real-world usage. To achieve this, we collected the data from multiple news websites under fair usage for educational and research purposes. The dataset collection process, preprocessing, statistics, and dataset quality are detailed in the following sections.

3.1.1 Data Collection

The data collection process for all eight Indian languages involved web scraping from multiple news websites. However, this task posed challenges due to the diverse and dynamic nature of these websites. Given that each website has its own unique structure, it was crucial to understand the intricacies of each site to extract data accurately, without any loss of information or introduction of noise. To achieve this, we developed site-specific web scrapers tailored to each website. These scrapers were designed to extract the text of news articles, headlines, and the name of the news subdomain. Care was taken to ensure that both the article and headline elements were non-empty and devoid of any unwanted information such as advertisements, URLs pointing to related articles, or embedded social media content.

To avoid any bias towards a particular news style, data was collected from a diverse range of news websites. Table 3.1 contains detailed list of websites used for scraping. These websites covered various

S.No	L	Website	S.No	L	Website	S.No	L	Website
1	te	https://www.ap7am.com/telugu-news	17	kn	https://kannadanewsnow.com/kannada/	33	ml	https://eveningkerala.com/
2	te	https://www.prabhanews.com/	18	kn	https://hosadigantha.com/	34	hi	https://www.jagran.com/
3	te	https://www.suryaa.com/index.html	19	kn	https://kannada.asianetnews.com/	35	hi	https://www.khaskhabar.com/
4	te	https://www.manatelangana.news/	20	kn	https://newskannada.com/	36	hi	https://www.indiatv.in/
5	te	http://www.andhrabhoomi.net/	21	kn	https://www.kannadaprabha.com/	37	bn	https://www.anandabazar.com/
6	te	https://prajasakti.com/	22	kn	https://www.sahilonline.net/ka	38	bn	https://www.sangbadpratidin.in/
7	te	https://www.vaartha.com/	23	kn	https://www.udayavani.com/	39	bn	https://bengali.abplive.com/live-tv
8	te	https://10tv.in/	24	kn	http://vishwavani.news/	40	bn	https://uttarbangasambad.com/
9	te	https://www.hmtvlive.com/	25	kn	https://ainlivenews.com/	41	bn	https://bangla.asianetnews.com/
10	ta	https://www.hindutamil.in/	26	kn	https://vaarte.com/	42	mr	https://www.lokmat.com/
11	ta	https://www.polimernews.com/	27	kn	https://btvkannada.com/	43	mr	https://prahaar.in/
12	ta	https://tamil.asianetnews.com/	28	ml	https://www.eastcoastdaily.com/	44	mr	https://marathi.abplive.com/
13	ta	https://www.updatenews360.com/	29	ml	https://suprabhaatham.com/	45	gu	https://sandesh.com/
14	kn	https://kannadadunia.com/	30	ml	https://www.bignewslive.com/	46	gu	https://www.gujaratsamachar.com/
15	kn	https://eesanje.com/	31	ml	https://www.malayalamexpress.in/	47	gu	https://gujarati.news18.com/
16	kn	https://www.vijayavani.net/	32	ml	https://dailyindianherald.com/			

Table 3.1: List of websites used for creating Mukhyansh.

domains, including state, national, international, entertainment, sports, business, politics, crime, and COVID-19, among others (refer Table 3.2). To ensure the quality of the collected data, additional preprocessing steps were implemented next.

3.1.2 Data Preprocessing

In the series of essential preprocessing steps, firstly, we eliminate all special symbols, emojis, and punctuation marks from the dataset. Next, we remove any duplicate article-headline pairs from the dataset. Lead or prefix, wherein the title of an article is derived from the initial sections that typically contain the most crucial information, is a widespread approach adopted by news sites. Although utilizing the lead section can be beneficial for summary generation, it may inadvertently hinder the model's ability to learn and discriminate between different types of information. By relying solely on the lead, the model may overlook relevant details and nuances present in the subsequent sections of the article. Therefore, we eliminate pairs with prefixes from the dataset. Furthermore, to ensure that only substantial and informative pairs are retained, we apply a minimum-length filter to the dataset. This filter helps eliminate article-headline pairs where the article contains fewer than 20 tokens and/or the headline consists of fewer than 3 tokens. Table 3.3 provides an overview of the preprocessing statistics for Mukhyansh and the final Train, Dev, and Test splits.

News Category	Category-wise counts of article-headline pairs for each language							
	te	ta	kn	ml	hi	bn	mr	gu
state	698059	133599	163857	144491	-	143804	184045	123183
national	91787	80711	61170	92833	314528	42913	72182	53248
entertainment	59244	31265	22697	14939	80202	31470	2819	19710
international	24262	29463	26092	34008	29668	20552	15347	37682
sports	19933	26186	18775	10204	78190	30676	29947	19337
business	13495	12874	8747	3446	60524	775	10379	21884
crime	8917	6656	7541	7064	8052	-	16489	-
covid	1425	6470	14147	4348	-	4205	-	-
politics	-	4484	5816	843	29459	346	3234	-
other	-	-	9081	2896	-	6532	-	914

Table 3.2: Category wise statistics of Mukhyansh

3.1.3 Data Splits

For the final splits, we allocated 90% of the data for training purposes, while the remaining data was dedicated to development and testing. To ensure robust performance and prevent any bias towards specific news categories or domains, stratified sampling techniques were employed when creating our data splits. This approach guarantees that articles from all categories are evenly distributed across the training, development, and test sets.

3.1.4 Data Statistics

To evaluate the task's abstractive nature and difficulty, we compute the percentage of novel n-grams and employ extractive baselines like LEAD-1 and EXT-ORACLE ROUGE-L (R-L) scores. The "percentage of novel n-grams" indicates the proportion of n-grams present in the headline but not found in the article, quantifying the level of uniqueness in the headline. Specifically, LEAD-1 R-L calculates the similarity between the first sentence of the article and the reference headline, while EXT-ORACLE R-L computes scores by selecting the sentence from the article that achieves the highest R-L scores with the reference headline. The resulting scores along with other statistics are detailed in Table 3.4

3.1.5 Human Evaluation

In order to evaluate the quality of the Mukhyansh dataset more comprehensively, a human evaluation was conducted. Due to resource constraints and the expenses associated with annotation, this evaluation was limited to the Telugu language data. A total of 500 article-headline pairs were randomly selected and assigned to native-language annotators. They were provided with a set of guidelines, which were based on those utilized in previous studies such as XL-Sum [14] and IndicNLG [18]. The evaluation specifically focused on the following properties:

- **Consistent** *True*, If the article and headline are consistent.
- **Inconsistent** *True*, If the headline contains information that is inconsistent with the article.
- **Unfounded** *True*, If the headline contains extra information that cannot be inferred from the article.

We assign each article-headline pair to 3 annotators and the final rating for each pair is selected based on majority voting. We found that 96.8% of the samples were rated *True* for *Consistency*, and the percentage of samples that are rated *Inconsistent*, and *Unfounded* were 0.6%, and 2.6% respectively, which supports our claim of a reliable and good-quality dataset.

The inter-annotator agreement was assessed using a variation of Fleiss' Kappa, proposed by [26] and it resulted in an encouragingly high score of 0.76, indicating substantial agreement among annotators.

	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 3.3: Statistics of Mukhyansh Preprocessing.

L	Total Pairs	Avg sents in article	Avg tokens in article	Avg tokens in headline	Total Tokens		Unique Tokens		% novel n-gram				Lead-1	EXT-ORACLE
					articles	headlines	articles	headlines	n=1	n=2	n=3	n=4	R-L	R-L
te	917122	7.97	103.64	7.42	95.05M	6.80M	2.3M	376K	36.63	62.87	82.10	91.41	23.54	33.21
ta	331708	15.47	218.99	11.50	72.64M	3.82M	1.8M	225K	33.02	55.12	73.75	85.05	32.70	39.33
kn	337923	10.94	154.77	9.03	52.3M	3.05M	1.9M	222K	41.30	65.88	82.73	91.45	19.66	30.08
ml	315072	10.26	115.45	9.54	36.37M	3.01M	2.5M	351K	36.14	55.59	71.20	81.73	34.60	41.94
hi	600623	14.54	303.05	13.45	182.02M	8.08M	1.3M	137K	20.31	47.20	67.96	81.27	25.99	35.02
bn	281273	19.41	244.78	10.10	68.85M	2.84M	0.9M	135K	37.60	67.60	84.31	92.27	15.51	30.50
mr	334442	17.71	271.02	8.41	90.64M	2.81M	1.9M	241K	37.11	64.73	82.66	91.66	13.88	28.34
gu	275958	16.45	284.39	12.46	78.48M	3.44M	1.7M	197K	38.24	65.81	82.08	90.54	12.21	28.72

Table 3.4: Mukhyansh dataset statistics in detail.

3.2 Baseline Models

In this work, we evaluate the performance of commonly used sequence-to-sequence models as baselines on our 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-trained transformer encoder-decoder models like mT5 [35] and IndicBART [6].

3.2.1 RNN based encoder-decoder network trained from scratch

Sequence to Sequence learning with recurrent neural networks [30] has had a significant impact on the field of NLP. Its ability to process sequential information, such as text and speech has enabled deep learning models to perform a wide range of text generation tasks like machine translation, and summarization with remarkable accuracy and paved the way for further advancements in NLP. It consists of two main components: an encoder and a decoder. The encoder component processes the input sequence and generates a context vector, which captures the important information in the input. The decoder component then uses this context vector to generate the output sequence. The encoder and decoder are typically implemented using Long Short-Term Memory (LSTM) [16] or Gated Recurrent Units (GRU) [4] networks. LSTMs and GRUs are popular choices due to their ability to effectively capture long-term dependencies in sequential data. This makes them well suited for encoding and decoding sequential inputs and outputs in the Sequence to Sequence model. However, these models had some limitations, One of the main limitations was that the fixed-length context vector could not effectively capture longer sequences, leading to the loss of information. Additionally, the decoder only had access to this context vector, which did not allow it to focus on different parts of the input sequence during the generation process. These limitations resulted in poor performance on longer sequences or tasks that required selective attention to different parts of the input. The attention mechanism was introduced in [1] to address these limitations by allowing the decoder to dynamically attend to different parts of the input sequence, providing a way to incorporate additional information and improving the model's performance.

In this subsection, we investigate different combinations of word embeddings and recurrent units used in our sequence-to-sequence models.

FastText Embeddings: FastText [11] is an enhanced version of the embedding model introduced by [21] in 2013. Operating on the principles of the skip-gram model, FastText incorporates subword information to create embeddings for individual words. Unlike traditional methods that directly learn word vectors, FastText represents each word as a collection of n-grams of characters. This approach enables FastText to effectively generate embeddings even for rare words.

Byte-Pair Embeddings (BPEmb): We leverage Byte Pair Encoding (BPE) [10] to address the challenges associated with word-level tokenization and embeddings, particularly prevalent in morphologically rich Indian languages. Models utilizing word-level tokenization and embeddings often encounter issues such as out-of-vocabulary (OOV) words and rare words. OOV words represent new terms not seen during training, resulting in inadequate embeddings, while rare words, infrequently appearing in training data, may receive weak or unreliable embeddings. BPE, an unsupervised learning method for subword-level tokenization, effectively mitigates these challenges by iteratively merging the most frequent pairs of bytes. This technique results in a subword vocabulary capable of capturing morphological information, thereby enhancing the robustness of the tokenization process. We use 300-dimensional subword embeddings from BPEmb [15], pretrained on Wikipedia using BPE technique.

3.2.1.1 FastText + GRU Model

- This model utilizes 300-dimensional pretrained FastText embeddings as the initial word representation.
- It employs GRU (Gated Recurrent Unit) networks in both the encoder and decoder.
- The model consists of 4 stacked layers, with each GRU cell containing 600 hidden activation units.

3.2.1.2 FastText + LSTM Model

- Similar to the previous model, this one also utilizes 300-dimensional pretrained FastText embeddings as the initial word representation.
- Instead of GRU, it utilizes LSTM (Long Short-Term Memory) networks in both the encoder and decoder.
- Like the previous model, it consists of 4 stacked layers, with each LSTM cell containing 600 hidden activation units.

3.2.1.3 BPEmb + GRU Model

- This model utilizes 300-dimensional pretrained Byte-Pair Embeddings (BPEmb) as the initial word representation.

- Similar to the previous models, it employs GRU networks in both the encoder and decoder.
- Like the other models, it consists of 4 stacked layers, with each GRU cell containing 600 hidden activation units.

3.2.2 Pre-trained transformer models

Pre-trained transformer models have been a revolutionary advancement in the field of Natural Language Processing (NLP). These models, built upon the innovative transformer architecture [32], have revolutionized how we approach language understanding tasks. By leveraging vast amounts of data, pre-training allows these models to learn intricate patterns and representations of language in an unsupervised or semi-supervised manner. This pre-training phase equips the models with a rich understanding of syntax, semantics, and context, making them highly versatile for a wide range of NLP tasks. Once pre-trained, these models can be fine-tuned on downstream tasks with task-specific labeled datasets. Fine-tuning involves updating the parameters of the pre-trained model using task-specific data, thereby adapting the model to perform well on the target task. Fine-tuning typically requires less data and computational resources compared to training models from scratch, making pre-trained transformer models highly versatile and efficient for various NLP applications.

We leverage the benefits of transfer learning in headline generation by utilizing pre-trained sequence-to-sequence models such as mT5 and IndicBART. To implement these models, we utilize the scripts[1] provided by Huggingface [34].

mT5: We conducted experiments on our dataset by fine-tuning the pre-trained mT5 [35] model, a multilingual variant of T5 [25], which was originally trained on the common crawl dataset, encompassing 101 languages. We use mT5-small model for our experiments.

IndicBART: IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on 11 Indian languages and English. It is similar to mBART [20] in terms of architecture and training methodology. Specifically, we use a variant of IndicBART called separate script IndicBART[2] (hereafter referred to as SSIB) and fine-tune it on our dataset for the task of headline generation.

3.2.3 Experimental Setup

3.2.3.1 Hyperparameters

In our experimental setup, we assessed various baseline models tailored to the task of headline generation for Indic languages, including Seq-Seq+FastText, Seq-Seq+BPEmb, mT5-small, and SSIB. Each model was configured with distinct parameters to optimize their performance. During the inference stage, our models employed a beam search strategy with a beam width of 5, except for mT5-small and SSIB, which utilized a beam width of 4. Beam search is a heuristic algorithm that explores a

[1] https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization
[2] https://huggingface.co/ai4bharat/IndicBARTSS

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 3.5: Experimental setup of various baseline models.

graph by expanding the most promising nodes within a limited set, determined by the beam width, at each level of the decoding process. To address the challenge of bias towards shorter sentences, length normalization was applied, with a length normalization penalty of 0.1 introduced for Telugu, Tamil, Kannada, and Malayalam languages, while no length normalization was applied to other languages. To prevent overfitting, we employ early stopping. Further details regarding the experimental setup and parameter configurations for all the models can be found in Table 3.5

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.

3.2.3.2 Metric

To assess the models' performance, we utilize the multilingual ROUGE metric [14][3], a widely used measure in NLP for assessing the quality of generated headline against reference headline. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) comprises several variants, with ROUGE-1, ROUGE-2, and ROUGE-L being the most commonly used ones.

ROUGE-1 evaluates the overlap of unigram tokens between the generated headline and the reference headline. It calculates precision, recall, and F1 score based on the number of overlapping unigrams, providing insight into the effectiveness of the model in capturing important content from the reference.

ROUGE-2 extends the evaluation to bigrams, measuring the overlap of consecutive pairs of words between the generated headline and the reference headline. Similar to ROUGE-1, it computes precision, recall, and F1 score based on the number of overlapping bigrams, providing a more nuanced evaluation by considering the continuity of phrases in the generated text.

ROUGE-L assesses the longest common subsequence (LCS) between the generated headline and the reference headline. It calculates precision, recall, and F1 score based on the length of the LCS, which

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

represents the maximum number of words that appear in both the generated and reference texts in the same order. ROUGE-L offers a more lenient evaluation compared to ROUGE-1 and ROUGE-2, as it considers word order but not necessarily word overlap.

3.2.4 Results

Table 3.6 presents the ROUGE-1, 2, L (R-1, R-2, R-L) scores achieved by different baseline models on Mukhyansh. The best R-L score for each language is highlighted in bold. Notably, the SSIB and mT5-small models outperformed all the sequence-to-sequence models trained from scratch. The superior performance of SSIB and mT5-small can be attributed to their pre-training on a large corpus.

It is worth mentioning that the GRU variant of the sequence-to-sequence model, utilizing FastText embeddings, yielded satisfactory results with a smaller parameter count (64 Million) compared to SSIB (244 Million) and mT5-small (300 Million). Examples of headlines generated by various baseline models fine-tuned on Mukhyansh presented in Figure 3.1 and Figure 3.2.

L	**FastText+GRU**			**FastText+LSTM**			**BPEmb+GRU**			**mT5-small**			**SSIB**		
	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 3.6: ROUGE-1,2,L scores of various baseline models of Mukhyansh for each language (L).

URL	https://www.jagran.com//news/national-five-children-killed-in-wall-collapse-incidents-10655913.html
Article	कौशांबी। उत्तर प्रदेश के कौशांबी जिले में दो अलग-अलग जगहों पर दिवार गिरने से पांच बच्चों की मौत हो गई। पुलिस ने सोमवार को बताया कि कौशांबी जिले के पथरावन गांव में रविवार शाम को मिट्टी से बने घर का दिवार अचानक गिर गया। दिवार के गिरने से सुभाष, उसकी बहन लक्ष्मी और कुंद्रा की दबने से मौत हो गई। पुलिस ने बताया कि दूसरी घटना अयाना एरिया के कारकापुर गांव में दिवार गिरने से दो बच्चे रजनीश और प्रियंका की भी मौत घटनास्थल पर ही हो गई। **Transliteration :** kauśāṃbī| uttara pradeśa ke kauśāṃbī jile meṃ do alaga-alaga jagahoṃ para divāra girane se pāṃca baccoṃ kī mauta ho gaī| pulisa ne somavāra ko batāyā ki kauśāṃbī jile ke patharāvana gāṃva meṃ ravivāra śāma ko miṭṭī se bane ghara kā divāra acānaka gira gayā। divāra ke girane se subhāṣa, usakī bahana lakṣmī aura kuṃdrā kī dabane se mauta ho gaī| pulisa ne batāyā ki dūsarī ghaṭanā ayānā eriyā ke kārakāpura gāṃva meṃ divāra girane se do bacce rajanīśa aura priyaṃkā kī bhī mauta ghaṭanāsthala para hī ho gaī। **Translation:** Kaushambi. In Uttar Pradesh's Kaushambi district, five children died due to wall collapse at two different places. Police said on Monday that the wall of a house made of mud suddenly collapsed in Pathravan village of Kaushambi district on Sunday evening. Subhash, his sister Lakshmi and Kundra died due to the collapse of the wall. Police said that in the second incident, two children Rajneesh and Priyanka also died on the spot due to wall collapse in Karkapur village of Ayana area.
Actual Headline	यूपी में दिवार गिरने से पांच बच्चों की मौत **Transliteration:** yūpī meṃ divāra girane se pāṃca baccoṃ kī mauta **Translation:** Five children died due to wall collapse in UP
GRU + FastText	यूपी में दो अलग जगहों पर दीवार गिरने से 5 बच्चों की मौत **Transliteration:** yūpī meṃ do alaga jagahoṃ para dīvāra girane se 5 baccoṃ kī mauta **Translation:** 5 children died due to wall collapse at two different places in UP
LSTM + FastText	यूपी में दो अलग हादसों से पांच की मौत **Transliteration:** yūpī meṃ do alaga hādasoṃ se pāṃca kī mauta **Translation:** Five killed in two separate accidents in UP
GRU + BPEmb	सड़क हादसे में 5 बच्चों की मौत **Transliteration:** saṛaka hādase meṃ 5 baccoṃ kī mauta **Translation:** 5 children died in road accident
mT5-small	उत्तर प्रदेश में दो जगहों पर दिवार गिरने से 5 बच्चों की मौत **Transliteration:** uttara pradeśa meṃ do jagahoṃ para divāra girane se 5 baccoṃ kī mauta **Translation:** 5 children died due to wall collapse at two places in Uttar Pradesh
SSIB	यूपी के कौशांबी में दो अलग अलग जगहों पर दिवार गिरने से 5 बच्चों की मौत **Transliteration:** yūpī ke kauśāṃbī meṃ do alaga alaga jagahoṃ para divāra girane se 5 baccoṃ kī mauta **Translation:** 5 children died due to wall collapse at two different places in UP's Kaushambi

Figure 3.1: Hindi example of headlines generated by various baseline models fine-tuned on Mukhyansh

URL	https://telangana.suryaa.com/telangana-updates-20874-.html
Article	బీజేపీ నేత బద్దం బాల్‌రెడ్డి మరణం తీరని లోటని అసెంబ్లి స్పీకర్ పోచారం శ్రీనివాసరెడ్డి అన్నారు. బద్దం బాల్‌రెడ్డి పార్థివదేహాన్ని సందర్శించి నివాళులర్పించారు. అనంతరం మాట్లాడుతూ ప్రజల మనిషిగా బాల్‌రెడ్డి గుర్తింపు తెచ్చుకున్నారన్నారు. హైదరాబాద్ ప్రజలతో బాల్‌రెడ్డికి అవినాభావ సంబంధం ఉందన్నారు. బాల్‌రెడ్డి కుటుంబ సభ్యులకు తన ప్రగాఢ సానుభూతి అన్నారు. **Transliteration :** bījepī neta baddaṃ bālrèḍḍi maraṇaṃ tīrani loṭani asèṃbli spīkar pocāraṃ śrīnivāsarèḍḍi annāru. baddaṃ bālrèḍḍi pārthivadehānni saṃdarśiṃci nivāḷularpiṃcāru. anaṃtaraṃ māṭlāḍutū prajala manişigā bālrèḍḍi gurtiṃpu tèccukunnārannāru. haidarābād prajalato bālrèḍḍiki avinābhāva saṃbaṃdhaṃ uṃdannāru. bālrèḍḍi kuṭuṃba sabhyulaku tana pragāḍha sānubhūti annāru. **Translation:** Assembly Speaker Pocharam Srinivas Reddy termed the death of BJP leader Baddam Bal Reddy as an irreparable loss. He visited the mortal remains of Baddam Bal Reddy and paid homage to him. Speaking after the meeting, he said that Bal Reddy has gained recognition as a people's man. Bal Reddy has a close relationship with the people of Hyderabad. My deepest condolences to the family members of Bal Reddy.
Actual Headline	బాల్‌రెడ్డి మరణం తీరని లోటు **Transliteration:** bālrèḍḍi maraṇaṃ tīrani loṭu **Translation:** Bal Reddy's death is an irreparable loss.
GRU + FastText	మృతి తీరని లోటు పోచారం **Transliteration:** mṛti tīrani loṭu pocāraṃ **Translation:** Death is an irreparable loss: Pocharam
LSTM + FastText	బీజేపీ నేత మృతి తీరని లోటు స్పీకర్ పోచారం **Transliteration:** bījepī neta mṛti tīrani loṭu spīkar pocāraṃṃ **Translation:** BJP leader's death is an irreparable loss: Speaker Pocharam
GRU + BPEmb	బీజేపీ నేత్త బద్దం బాల్రెడ్డి మృతి తీరనిలోటు స్పీకర్ పోచారం శ్రీనివాసరెడ్డికి సంతాపాలర్పించిన అసెంబ్లి స్పీకర్ **Transliteration:** bījepī neta baddaṃ bālrèḍḍi mṛti tīraniloṭu spīkar pocāraṃ śrīnivāsarèḍḍiki saṃtāularpiṃcina asèṃbli spīkar **Translation:** BJP leader Baddam Bal Reddy's death is a sad loss: Assembly Speaker pays condolences to Speaker Pocharam Srinivasa Reddy
mT5-small	బద్దం బాల్ రెడ్డి మరణం తీరని లోటు పోచారం **Transliteration:** baddaṃ bāl rèḍḍi maraṇaṃ tīrani loṭu pocāraṃ **Translation:** Baddam Bal Reddy's death is an irreparable loss: Pocharam
SSIB	బద్దం బాల్ రెడ్డి మరణం తీరని లోటు స్పీకర్ పోచారం **Transliteration:** baddaṃ bāl rèḍḍi maraṇaṃ tīrani loṭu spīkar pocāraṃ **Translation:** Baddam Bal Reddy's death is an irreparable loss: Speaker Pocharam

Figure 3.2: Telugu example of headlines generated by various baseline models fine-tuned on Mukhyansh

3.3 Existing datasets evaluation

Due to the unavailability of publicly accessible data from existing monolingual works, our evaluation is limited to the recent multilingual datasets, namely XL-Sum and IndicHG. The Indian language section of the XL-Sum dataset consists of 251K article-headline pairs sourced from BBC[4]. While XL-Sum focuses on extreme summarization, it is important to note that the summaries provided may consist of more than one sentence. Additionally, concerns have been raised by [31] regarding the quality of summaries in the Indian language section of XL-Sum. Consequently, our evaluation is primarily centered on the IndicHG dataset[5].

To validate the reported results in IndicNLG regarding headline generation, we conduct a series of experiments on the IndicHG dataset, accompanied by comprehensive quantitative and qualitative analyses. As discussed in the subsequent sub-sections, our investigation has uncovered significant quality issues with the HG dataset of IndicNLG. Despite the valuable contributions of IndicNLG to the field of language generation for various Indic languages, it is imperative to address these issues before deeming the IndicHG dataset suitable for training robust models.

3.3.1 Reproducing IndicHG Results

We initiate our experiments with an attempt to replicate the findings of IndicHG for the eight Indian languages mentioned. Following their paper's methodology and hyper-parameter settings, we meticulously fine-tune the SSIB model, (hereafter, referred to as *IndicHG**). In order to obtain a more reliable assessment of the model's performance and evaluate the consistency of the results, we conducted the same experiment five times with different initial seeds. Subsequently, In Table 3.7 we present the mean and standard deviation of the ROUGE-1,2,and L scores obtained on the test set. Table 3.8 presents these mean ROUGE-L scores alongside their reported[6] counterparts.

As depicted in the final row of Table 3.8, there is an average reduction of 17.85 in the ROUGE-L scores across the eight languages. This substantial decrease raises concerns regarding the reproducibility of the original findings and emphasizes the necessity for further investigation.

3.3.2 Quantitative Analysis

We initiate the analysis by implementing preprocessing steps for the IndicHG dataset, including checks for prefixes, duplicates, and minimum length. In addition to the eight languages we are focusing on, we extended the preprocessing to include the remaining three languages of IndicHG: Oriya, Punjabi, and Assamese.

[4] `https://www.bbc.com/`

[5] IndicNLG data for Headline-generation was taken from `https://huggingface.co/datasets/ai4bharat/IndicHeadlineGeneration/tree/main/data`

[6] The reported scores are taken from the monolingual works of IndicHG [18] paper, as the checkpoint is not made public.

L	R-1		R-2		R-L	
	mean	std	mean	std	mean	std
te	23.75	1.31	11.98	0.88	22.37	1.28
ta	34.49	0.70	21.06	0.62	32.96	0.74
kn	43.85	1.41	35.89	1.58	42.79	1.43
ml	37.20	1.62	25.59	1.89	35.64	1.72
hi	28.73	0.63	13.42	0.37	24.12	0.62
bn	24.54	0.29	12.58	0.36	22.54	0.31
mr	22.99	0.46	11.26	0.28	21.28	0.38
gu	24.77	0.22	11.87	0.15	22.68	0.35
Average	30.04	0.83	17.96	0.77	28.05	0.85

Table 3.7: Mean & Standard Deviation of 5 iterations of IndicHG* results

Surprisingly, despite claims to the contrary, 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 3.9 demonstrate a high level of overlap among these splits for most of the languages, corroborating data contamination. Figure 3.3 shows the duplicates and contaminated(overlap with train/dev set) article-headline pairs percentage in IndicHG test-set. For instance, an article-headline pair[7] from the Kannada language appears 115 times in the training data, 18 times in the development data, and 2 times in the test data. Figure 3.4 depicts the percentages of duplication remained in train, dev and test splits of IndicHG after removing all overlapping pairs.

Data contamination introduces bias in evaluation, as the metrics calculated on the development and test datasets do not accurately represent the model's performance on unseen data. Additionally, we assert that the heavy presence of duplicated data in the dataset may lead models trained on this data to achieve artificially high performance by memorizing the duplicated pairs, thereby hindering their ability to generalize to new, unseen data.

To support our arguments, we take several steps. Firstly, we eliminate all duplicate pairs from each of the training, development, and test splits of the IndicHG dataset. To deal with data contamination, the following 2 variations were attempted:

1. To ensure the integrity of the test set, a straightforward approach was adopted, which involved excluding any pairs that were already present in the corresponding train/dev sets. Additionally, any

[7] https://tinyurl.com/2p85mayt

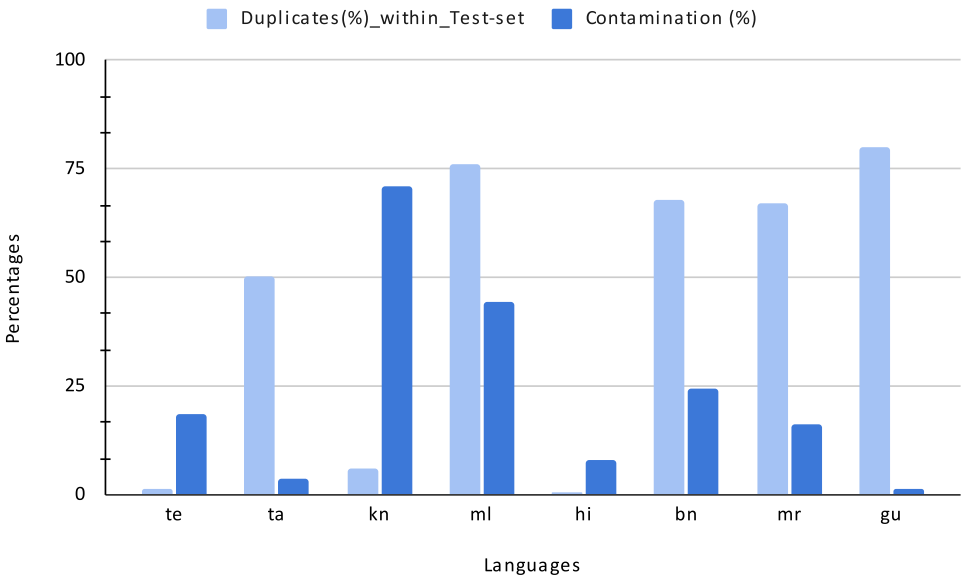

Figure 3.3: Language-wise data bias in IndicHG test-set.

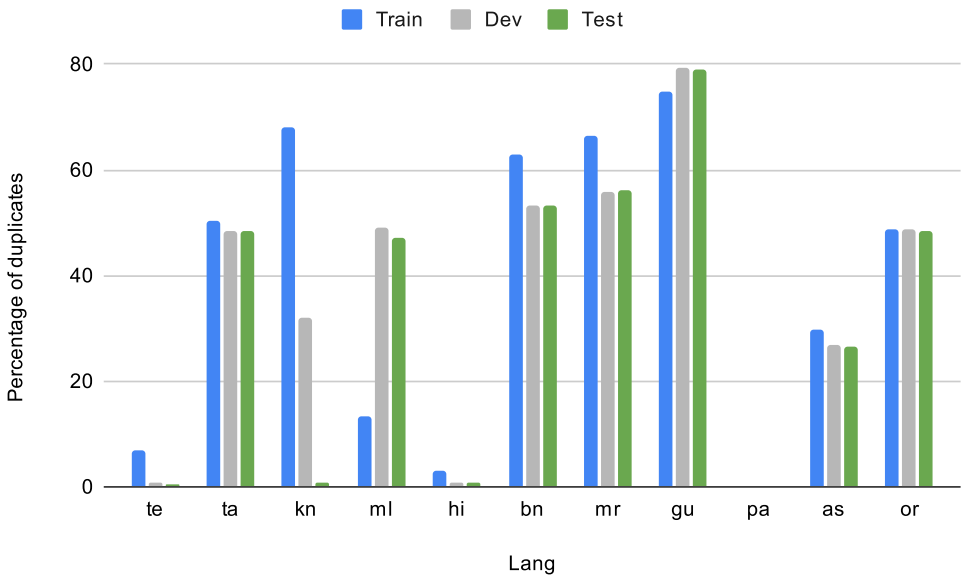

Figure 3.4: Duplication percentage within IndicHG train,dev,test splits.

IndicHG Performance			
L	**Reported**	**Reproduced**	**Unbiased**
te	41.97	22.37	19.47
ta	46.52	32.96	33.79
kn	73.19	42.79	21.64
ml	60.51	35.64	26.79
hi	34.49	24.12	22.68
bn	37.95	22.54	20.28
mr	40.78	21.28	20.14
gu	31.80	22.68	22.61
Average	45.90	28.05	23.42
Performance drop		17.85	22.48

Table 3.8: Performance Comparison of various versions of IndicHG: Reported, IndicHG* and IndicHG_Unbiased.

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 3.9: IndicHG Analysis: Showing overall duplication and overlap(or data-contamination) percentages.

pairs in the dev set that were already present in the train set were also removed. This approach effectively eliminated data contamination and allowed the training set to remain as large as possible.

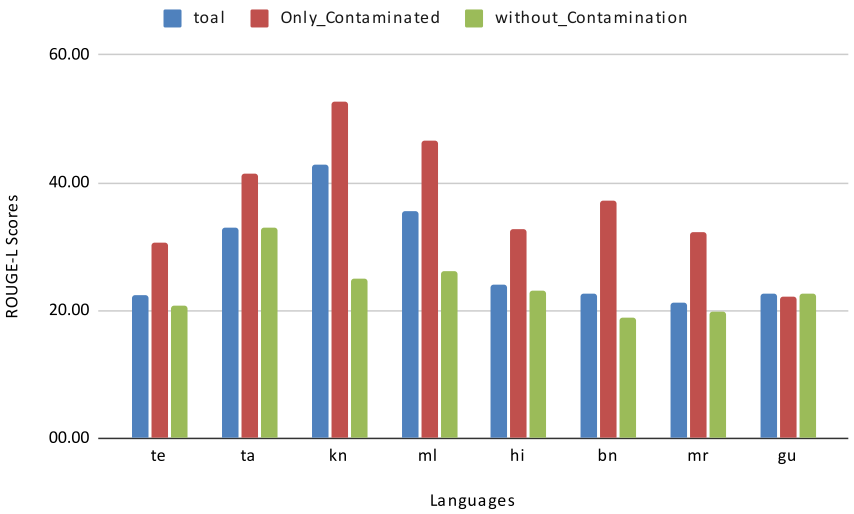

Figure 3.5: ROUGE-L scores for subsets of IndicHG Test set.

These splits were then utilized to reproduce the IndicHG results as *IndicHG_Unbiased*. Notably, this dataset exhibited a significant decrease in average R-L score, with a decrease of 22.48 compared to the score reported in the original IndicNLG paper [18], resulting in an average R-L score of 23.42; as outlined in Table 3.8.

To evaluate the specific impact of data contamination, we divided the IndicHG test set into two subsets. The first subset consisted of pairs from the IndicHG test set that were also present in the corresponding train or dev sets. The second subset comprised the remaining (unique) pairs from the original test set. Figure 3.5 and Table 3.10 shows the R-L score comparison for these two test subsets, referred to as *Overlaps* and *Without_Overlap* respectively, against those obtained from the total (original) test set. The results unequivocally support the claim that data contamination indeed leads to artificial high performance.

2. As an alternative approach, pairs present in the training set that also appeared in the corresponding dev and test sets were eliminated. Similarly, pairs in the dev set that were already present in the test set were excluded. Additionally, pairs were filtered out if the headline was found in the article's prefix, or if the pairs were too short. This method aimed to ensure that the new test set closely resembled the original set while eliminating problematic cases. The stepwise statistics of this filteration process and final split counts are provided in Table 3.11. Further statistics of the resulting filtered dataset, referred to as *IndicHG_filtered*, can be found in Table 3.12.

While it may seem intuitive that a larger training set would lead to better model training, our findings suggest that both of the aforementioned approaches yield similar scores. Consequently, we have decided to utilize the *IndicHG_filtered* version for all future cross-comparisons. This is primarily because its

Test sets	Models Fine-tuned on	Language								Average
		te	ta	kn	ml	hi	bn	mr	gu	
IndicHG	IndicHG*	22.37	32.96	42.79	35.64	24.12	22.54	21.28	22.68	28.05
Overlaps (IndicHG)		30.53	41.36	52.63	46.61	32.81	37.12	32.21	22.27	36.94
IndicHG-Overlaps		20.84	32.87	25.00	26.07	23.08	18.79	19.92	22.53	23.64

Table 3.10: Impact of Overlap on IndicHG Performance (by ROUGE-L).

	Dravidian language family				Indo-Aryan language family			
	te	ta	kn	ml	hi	bn	mr	gu
Total # Pairs	26717	75954	155057	20966	297284	142731	142590	262457
# Duplicates	2772	38958	135693	11276	13638	100835	99433	200334
# Pairs after deduplication	23945	36996	19364	9690	283646	41896	43157	62123
# Pairs with prefix	669	796	5	22	19	336	6	92
# Pairs with multiple-articles	5773	0	0	0	0	0	0	0
# Pairs too short	30	5	7	8	1470	5	8	7
# Pairs after filtering	17473	36195	19352	9660	282157	41555	43143	62024
# Pairs in train	13539	28750	13602	7235	194627	32435	33772	49566
# Pairs in dev	1903	3702	2693	1177	43604	4480	4644	6228
# Pairs in test	2031	3743	3057	1248	43926	4640	4727	6230

Table 3.11: IndicHG_filtered dataset creation statistics.

L	Total Pairs	Avg sents in article	Avg tokens in article	Avg tokens in headline	Total Tokens		Unique Tokens		% novel n-gram				Lead-1	EXT-ORACLE
					articles	titles	articles	titles	n=1	n=2	n=3	n=4	R-L	R-L
te	17473	13.99	185.09	7.97	3.2M	139K	238K	35.6K	36.26	65.66	85.87	94.08	15.23	29.30
ta	36195	13.43	181.77	11.76	6.58M	425K	311K	51.9K	32.94	54.89	70.17	78.96	33.46	40.10
kn	19352	11.49	189.16	9.22	3.66M	178K	237K	34.1K	33.02	57.47	75.89	86.59	18.19	29.73
ml	9660	13.55	168.38	10.08	1.63M	97K	232K	29K	39.15	61.48	77.78	87.04	26.40	35.79
hi	282157	18.25	397.08	12.55	112.04M	3.5M	543K	74.9K	20.79	49.86	71.06	83.56	21.99	32.52
bn	41555	14.65	239.55	11.27	9.95M	468K	245K	47.2K	38.02	64.35	80.91	89.33	13.95	27.39
mr	43143	13.61	205.31	8.57	8.86M	369K	258K	51K	31.45	57.76	77.44	86.59	13.08	32.55
gu	62024	12.31	226.64	11.20	14.06M	694K	425K	81.5K	35.85	60.69	76.75	85.94	15.52	29.39

Table 3.12: IndicHG_filtered dataset statistics in detail.

test set bears closer resemblance to the original test set. Section 3.3.4 describes further experimentation conducted using this dataset.

3.3.3 Qualitative Analysis:

To conduct a qualitative analysis, we begin by manually evaluating a random selection of article-headline pairs from the IndicHG Telugu dataset[8]. This dataset comprises articles collected from approximately 22 different Telugu news websites. To ensure a comprehensive evaluation, we assess at least five random pairs from each website. Our evaluation brings to light certain issues that indicate a lack of site-specific scraping implementation in IndicHG. The identified issues are as follows:

1. Unwanted information (noise) is present at the beginning of the article (Figure 3.6).
2. Headline is out of the context of the article (Figure 3.7).
3. The article part of a pair, itself contains multiple other article-headline pairs (Figure 3.8).

These quality issues in the article-headline pairs can significantly impact the performance of models. When the headline is contextually unrelated to the article, the generated headlines by the model are inaccurate, resulting in subpar performance. Likewise, the presence of multiple articles within a single article introduces irrelevant information, causing the model to focus on only a fraction of the total content.

For each of the aforementioned issues, we meticulously document the corresponding source website. Subsequently, we employ simple scripts, regular expressions, and other techniques to further examine all the article-headline pairs from these source websites. Among all the issues observed, the most prevalent is the occurrence of multiple articles within a single article (issue-3). By employing basic regular expressions, we were able to detect a total of 5773 such pairs, although not capturing all instances, primarily sourced from the Andhra Bhoomi website[9], which constitutes 30% of the Telugu IndicHG dataset. Considering the significant quantity of such pairs, we further update our *IndicHG_filtered* dataset by eliminating these pairs. The analysis of article-headline pairs of BBC Telugu[10], BBC Tamil[11] websites that are present in IndicHG dataset is detailed in Table 3.13.

3.3.4 Experiments and Analysis

In order to assess the effectiveness of different models, we fine-tune the SSIB model[12] using a range of specifically crafted training and test sets:

1. First, we fine-tune a model on the IndicHG_filtered dataset and evaluate its performance on the corresponding filtered test set, while ensuring that the fine-tuning hyperparameters remain consistent with those described in the IndicNLG paper. The results, as presented in Table 3.14, demonstrate the true performance of IndicHG when only good quality unique pairs are considered.

[8] Manual evaluation was restricted to Telugu, due to limited language experts/resources.
[9] http://www.andhrabhoomi.net/
[10] https://www.bbc.com/telugu
[11] https://www.bbc.com/tamil
[12] Unless otherwise stated, all experiments conducted in this study were based on the SSIB model.

URL	https://www.bbc.com/telugu/india-48363611
Article	వైసీపీ మెజారిటీకి ప్రజాశాంతి పార్టీ గండికొట్టిందా? ఒకే పేరుతో నిలబెట్టిన అభ్యర్థులకు వచ్చిన ఓట్లెన్ని? 24 మే 2019 దీనిని క్రింది వాటితో షేర్ చేయండి ఇవి బయటి లింక్లు, కాబట్టి కొత్త విండోలో తెరవబడతాయి ఇవి బయటి లింక్లు, కాబట్టి కొత్త విండోలో తెరవబడతాయి షేర్ ప్యానెల్ను మూసివేయండి ఆంధ్రప్రదేశ్ ఎన్నికల్లో కేఏ పాల్ నేతృత్వంలోని ప్రజాశాంతి పార్టీ చాలా చోట్ల తన అభ్యర్థులను బరిలోకి దింపింది.కొన్ని చోట్ల వైసీపీ అభ్యర్థుల పేర్లను పోలిన వ్యక్తులను బరిలోకి దింపిందనే వార్తలు వచ్చాయి.దీనిపై వైసీపీ ప్రతినిధులు మార్చి 26న దిల్లీకి వచ్చి ఎన్నికల సంఘానికి ఫిర్యాదు కూడా చేశారు.దాదాపు 35 నియోజకవర్గాల్లో తమ అభ్యర్థులను పోలిన అభ్యర్థులను ప్రజాశాంతి పోటీలో నిలబెట్టిందని, దీనిపై చర్యలు తీసుకోవాలని కోరింది.ప్రజాశాంతి ఎన్నికల గుర్తు అయిన హెలికాప్టర్ కూడా తమ ఫ్యాన్ గుర్తును పోలి ఉందని, దీనిపైనా చర్యలు తీసుకోవాలని కోరింది.అయితే, కేఏ పాల్ నిలబెట్టిన అభ్యర్థుల వల్ల వైసీపీకి నష్టం జరిగిందా..? ఏ నియోజకవర్గాల్లో వైసీపీ అభ్యర్థుల మెజారిటీపై ప్రభావం పడింది? ఫలితాలు ఎలా ఉన్నాయి? అనేది కింది పట్టికలో చూడొచ్చు.క్రమసంఖ్య
Translation	Did Prajashanti Party affect the YCP majority? How many votes did the candidates with the same name get? 24 May 2019 Share this with the following These are external links, so will open in a new window These are external links, so will open in a new window Close share panel In the Andhra Pradesh elections, Prajashanti Party led by KA Paul has fielded its candidates in many places. In some places, there were reports that Prajashanti candidates names are similar to YCP candidates. For this, YCP representatives came to Delhi on March 26 and filed a complaint to the Election Commission. Praja Shanti has fielded candidates similar to their candidates in about 35 constituencies and asked them to take action on this. Praja Shanti's symbol of the election, the helicopter, is also similar to their fan symbol, and they have asked for action on this too. However, has the YCP lost because of the candidates fielded by KA Paul..? In which constituencies has the majority of YCP candidates been affected? How are the results? can be seen in the following table. Serial no
Headline	వైసీపీ మెజారిటీకి ప్రజాశాంతి పార్టీ గండికొట్టిందా? ఒకే పేరుతో నిలబెట్టిన అభ్యర్థులకు వచ్చిన ఓట్లెన్ని? – BBC News తెలుగు
Translation	Did Prajashanti Party affect the YCP majority? How many votes did the candidates with the same name get? - BBC News Telugu
Explanation	The text highlighted in cyan color is the prefix information which is the same as the headline, and the one in pink is unwanted information (noise).

Figure 3.6: Example of Prefix Case

URL	https://www.bbc.com/telugu/india-42493669
Article	ఇన్స్టెంట్ ట్రిపుల్ తలాక్కు చెళ్లు దశాబ్దాలుగా ఎంతో మంది ముస్లిం మహిళల వేదనకు కారణమైన విధానం..'ఇన్స్టెంట్ ట్రిపుల్ తలాక్'.ఈ ఇస్లామిక్ ఆచారాన్ని రాజ్యాంగ విరుద్ధమని తీర్మానిస్తూ దేశ అత్యున్నత న్యాయస్థానం ఆగస్టులో చారిత్రక తీర్పుని వెలువరించింది.ఐదుగురు సభ్యులున్న ధర్మాసనంలో ముగ్గురు జడ్జిలు 'ఇన్స్టెంట్ ట్రిపుల్ తలాక్' రాజ్యాంగ విరుద్ధమనీ , అది మహిళలపై వివక్ష చూపేదిగా ఉందనీ పేర్కొన్నారు.సుప్రీం కోర్టు ప్రకటించిన ఈ నిర్ణయం పట్ల దేశ ప్రజలు, ముఖ్యంగా ముస్లిం మహిళలు హర్షం వ్యక్తం చేశారు.పార్లమెంటులో ట్రిపుల్ తలాక్ బిల్లుని సైతం ప్రవేశ పెట్టడంతో దానికి సంబంధించిన చట్ట రూపకల్పనలో మరో ముందడుగు పడింది. కానీ కొన్ని ముస్లిం మహిళా సంఘాలు, ఆల్ ఇండియా ముస్లిం పర్సనల్ లా బోర్డ్తో సహా కొన్ని రాజకీయ పార్టీలు మాత్రం ఆ బిల్లుని వ్యతిరేకిస్తున్నాయి. భారత్ నం.1 ఎక్కువ మంది భారతీయులు ఇష్టపడే క్రీడ క్రికెట్.కోహ్లీ సిక్స్ కొట్టినా, బుమ్రా వికెట్ తీసినా అది తమ ఘనతేనన్నట్టు క్రీడాభిమానులు సంబర పడతారు.అలాంటి అభిమానులను ఉత్సాహ పరిచే మరో అరుదైన మైలురాయిని భారత క్రికెట్ జట్టు ఈ ఏడాది తొలిసారి నమోదు చేసింది.సెప్టెంబరులో ప్రకటించిన ఐసీసీ ర్యాంకుల్లో ఆటు టెస్టులూ, ఇటు వన్డేల్లోనూ భారత్ నంబర్. 1 స్థానాన్ని కైవసం చేసుకుంది.ఒకేసారి ఇలా రెండు ఫార్మాట్లలో తొలి స్థానంలో నిలవడం భారత జట్టుకి ఇదే మొదటిసారి.కాగా, రోహిత్ శర్మ ఈ ఏడాది చివరలో వన్డేల్లో మూడో ద్విశతకం సాధించి ప్రపంచ రికార్డు నెలకొల్పాడు.మిథాలీ సారథ్యంలో భారత మహిళా క్రికెట్ జట్టు ప్రపంచ కప్ ఫైనల్కు చేరి రన్నరప్గా నిలిచింది.ఫిబ్రవరిలో జరిగిన '2017 బ్లైండ్ వరల్డ్ టీ20' క్రికెట్ టోర్నీని కూడా భారత అంధుల క్రికెట్ జట్టే గెలుచుకుంది.మరోపక్క బ్యాడ్మింటన్లో తెలుగు కుర్రాడు కిదాంబి శ్రీకాంత్ కొత్త చరిత్ర సృష్టించాడు.ఒక ఏడాదిలో నాలుగు సూపర్ సిరీస్ టైటిళ్లు గెలుచుకున్న తొలి భారతీయుడిగా రికార్డు నెలకొల్పాడు.ఇండొనేసియా, ఆస్ట్రేలియా, డెన్మార్క్, ఫ్రాన్స్ దేశాల్లో జరిగిన సూపర్ సిరీస్ టోర్నీల్లో శ్రీకాంత్ విజేతగా నిలిచాడు.
Translation	Ban Instant triple talaq Instant triple talaq is a practice that has caused the agony of many Muslim women for decades. The Supreme Court of the country issued a historic verdict in August declaring this Islamic practice unconstitutional. Three judges in a five-member bench ruled that instant triple talaq is unconstitutional and discriminatory against women. The people of the country, especially the Muslim women, were happy about the decision announced by the Supreme Court. Another step forward in the drafting of the law was made with the introduction of the Triple Talaq Bill in the Parliament. But some Muslim women's groups and some political parties, including the All India Muslim Personal Law Board, are opposing the bill. Bharat number one Cricket is India's No. 1 favorite sport of most Indians. Even if Kohli hits a six or Bumrah takes a wicket, the sports fans celebrate as if it is their honor. Another rare milestone that excites such fans has been recorded by the Indian cricket team for the first time this year. In the ICC rankings announced in September, both Tests and ODIs India no. 1 position. This is the first time for the Indian team to be ranked first in two formats at the same time. Meanwhile, Rohit Sharma set a world record by scoring his third double century in ODIs at the end of this year. Under the leadership of Mithali, the Indian women's cricket team reached the final of the World Cup and became the runner-up. The '2017 Blind' held in February The Indian blind cricket team also won the World T20 cricket tournament. On the other hand, Telugu boy Kidambi Srikanth created a new history in badminton. He set a record as the first Indian to win four super series titles in one year. Srikanth became the winner in the super series tournaments held in Indonesia, Australia, Denmark and France.
Headline Translation	దంగల్ బాహుబలి.. రెండూ రెండే – BBC News తెలుగు Dangal Baahubali.. both are .. - BBC News Telugu
Explanation	The text highlighted in cyan is the headline of the article (highlighted in lime), and the text highlighted in yellow is the headline of the article (highlighted in gray). Here, the actual headline has no context in the article.

Figure 3.7: Example of headline out of context from article

URL	http://www.andhrabhoomi.net/content/dudddd
Article	ముంబయి: దుబాయి నుంచి ఇక్కడికి వచ్చిన విమానంలో పోలీసులు సోదాలు చేయగా టాయిలెట్లో 3 కిలోల బంగారం బయట పడింది.దుబాయి నుంచి వచ్చిన ప్రయాణికుల్లో ఎవరో ఈ బంగారాన్ని తెచ్చి టాయిలెట్లో వదిలేసి ఉంటారని పోలీసులు చెబుతున్నారు.కస్టమ్స్ తనిఖీల్లో దొరికిపోతే కేసులు పెడతారన్న భయంతో ఇలా బంగారాన్ని వదిలేసి ఉంటారని పోలీసులు అనుమానిస్తున్నారు. భారత్ ఎదుగుదలలో యూపీ కీలకం లక్నో: భారత్ అయిదు ట్రిలియన్ డాలర్ల ఆర్థిక వ్యవస్థగా అవతరించడంలో, 2030 నాటికి ప్రపంచంలోని మూడు అత్యంత పెద్ద ఆర్థిక వ్యవస్థలలో ఒకటిగా ఎదగడంలో ఉత్తర్ప్రదేశ్ ఒక ముఖ్యమయిన పాత్ర పోషిస్తుందని రక్షణ శాఖ మంత్రి రాజ్నాథ్ సింగ్ అన్నారు. పది మందికి కేబినెట్ పదవులు బెంగళూరు, ఫిబ్రవరి 6: కర్నాటకలో కాంగ్రెస్– జేడీఎస్ సంకీర్ణ ప్రభుత్వాన్ని కుప్పకూల్చి బీజేపీ అధికారంలోకి రావడానికి సహకరించిన 10 మంది ఫిరాయింపు దారులకు ముఖ్యమంత్రి యెడ్యూరప్ప మంత్రి వర్గంలో కేబినెట్ పదవులు లభించాయి. 'ఇంటర్నెట్' ప్రాథమిక హక్కుకాదు న్యూఢిల్లీ, ఫిబ్రవరి 6: ఇంటర్నెట్ వినియోగించుకునే హక్కు ప్రాథమిక హక్కు కాదని, అది ఎంత మాత్రం దేశ భద్రతతో సమానమైన ప్రాధాన్యతను కలిగి ఉన్నది కాదని కేంద్ర మంత్రి రవిశంకర్ ప్రసాద్ గురువారం రాజ్యసభలో ప్రకటన చేశారు.దేశ భద్రతా పరిస్థితులను కూడా అంతే ప్రాధాన్యతతో పరిశీలించాల్సిన అవసరం ఉందన్నారు.
Translation	MUMBAI: When the police searched the flight that came here from Dubai, 3 kg of gold was found in the toilet. The police say that some of the passengers from Dubai must have brought this gold and left it in the toilet. UP is crucial for India's growth Defense Minister Rajnath Singh has said that Uttar Pradesh will play an important role in India becoming a five trillion dollar economy and one of the world's three largest economies by 2030. Cabinet posts for 10 people Bengaluru, Feb 6: The Congress-JD(S) coalition government in Karnataka was brought down by the BJP and helped to bring it to power. 10 defectors got cabinet posts in Chief Minister Yeddyurappa's cabinet. 'Internet' is not a fundamental right New Delhi, February 6: Union Minister Ravi Shankar Prasad on Thursday announced in the Rajya Sabha that the right to use the Internet is not a fundamental right and it does not have the same priority as national security. He said that there is a need to examine the situation with the same priority.
Headline Translation	విమానం టాయిలెట్లో 3 కిలోల బంగారం స్వాధీనం 3 kg gold seized in plane toilet
Explanation	Example of article-headline pair with multiple unrelated articles and headlines present in the same piece of text. The text highlighted in cyan color is the headline, followed by its article highlighted in yellow.

Figure 3.8: Multiple article-headline pairs in single article text

Error Cases	Telugu	Tamil
# Pairs	1587	3800
# Pairs with headline present in prefix	484	1558
# Pairs with unwanted information in the article	1390	3494
# Pairs with above two issues in common	461	1436
# Pairs with headline that is out of the context to the article	174	184

Table 3.13: Statistics of problematic pairs of IndicHG dataset (BBC Website).

It is evident that the ROUGE-L scores decrease significantly compared to the scores produced by the biased data (i.e. unfiltered IndicHG). Next, other models were also tested on IndicHG_filtered test set. See, Table 3.14. Notably, while testing IndicHG* model on IndicHG_filtered test set, we are bound to get biased (high) scores. This is because in case of IndicHG_filtered, the training set itself was prepared without overlapping pairs (leaving them intact in the corresponding test set). Keeping this bias aside, our Mukhyansh model outperforms all the others.

2. To further investigate the impact of quality vs. quantity, we prepare a smaller version of the Mukhyansh dataset. In order to create the new train, dev, and test sets, separate random sampling is performed over the original train, dev, and test sets of Mukhyansh. A model, called *Mukhyansh_small*, is then fine-tuned only on this smaller train set, and tested against other models, see Table 3.14.

This cross-comparison was then concluded by testing Mukhyansh's SSIB baseline against all other test sets. And as evident by the R-L scores (highlighted as bold) in Table 3.14 Mukhyansh outperforms almost all the other models.

Test sets	Models Fine-tuned on	Language								Average
		te	ta	kn	ml	hi	bn	mr	gu	
IndicHG	IndicHG_filtered	17.80	33.46	22.98	25.09	24.52	19.18	21.35	22.87	23.41
	Mukhyansh	**27.05**	**35.05**	**28.20**	**29.23**	**26.84**	17.65	**26.31**	19.86	**26.27**
	Mukhyansh_small	21.46	30.68	23.09	24.04	23.39	15.66	22.22	18.59	22.39
IndicHG_filtered	IndicHG* (with overlap)	20.58	32.50	41.89	33.29	23.92	21.78	21.93	22.51	27.30
	IndicHG_filtered	16.66	32.85	22.91	25.11	24.61	18.97	21.38	22.86	23.17
	Mukhyansh	**24.00**	**34.96**	**27.98**	**29.28**	**26.95**	17.66	**26.33**	19.85	**25.88**
	Mukhyansh_small	19.67	30.54	22.98	24.00	23.50	15.66	22.24	18.59	22.15
Mukhyansh	IndicHG*	19.83	29.31	20.61	19.51	23.95	14.80	15.29	16.19	19.94
	IndicHG_filtered	17.53	29.43	18.66	20.44	26.07	16.13	15.73	16.56	20.07
	Mukhyansh	**37.33**	**41.16**	**32.59**	**32.04**	**36.18**	**22.04**	**27.08**	**23.05**	**31.43**
	Mukhyansh_small	28.66	36.01	26.11	26.73	32.39	18.96	22.59	20.49	26.49

Table 3.14: Performance comparison (by ROUGE-L) of various models.

Chapter 4

Relevance-based Headline Generation for Telugu

Headline generation models are commonly trained on datasets obtained through web scraping. If the dataset contains a mixture of relevant headlines, clickbait, sensational, and misleading, it can bias the model towards generating similar types of headlines. Therefore, it is crucial to have high-quality and relevant training data to train headline generation models capable of producing highly relevant headlines. With this motivation, we developed a novel dataset, which captures the intricacies of real-world news article-headline pairs in the Telugu language. This chapter explores the creation of "TeClass", a human annotated relevance-based headline classification dataset for Telugu. This chapter also presents how different categories of news headlines impact the performance of headline generation models through experiments.

4.1 TeClass Dataset

4.1.1 Source

We utilize a subset of article-headline pairs from the Mukhyansh dataset, which covers a diverse set of news domains and websites. To further enhance the diversity of the dataset, we collect article-headline pairs from two additional websites, namely filmibeat[1] and gulte[2].

4.1.2 Annotation

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

[1] https://telugu.filmibeat.com/
[2] https://telugu.gulte.com/

volunteers. We presented the following instructions to the annotators, and 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.

Highly Related (HREL): The headline is highly related to the article content if it satisfies the following condition (Example presented in Figure 4.1):

- 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.

Moderately Related (MREL): The headline is moderately related to the article content if it satisfies any of the following conditions (Example presented in Figure 4.2):

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

Least Related (LREL): The headline is least related to the article content if it satisfies any of the following conditions (Example presented in Figure 4.3):

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

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

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 4.1: Example of Highly Related Headline

Article	అమరావతి : రెండు తెలుగు రాష్ట్రాల మధ్య జల వివాదం ఏర్పడిన నేపథ్యంలో కృష్ణా, గోదావరి నదీ జలాల బోర్డుల పరిధులను ఖరారుచేస్తూ మొన్న అర్ధరాత్రి కేంద్ర జలశక్తి మంత్రిత్వ శాఖ గెజిట్ను విడుదల చేసిన విషయం తెలిసిందే. దీనిపై టీడీపీ అధినేత చంద్రబాబు నాయుడు స్పందించారు. ఆ గెజిట్పై పూర్తిగా అధ్యయనం చేశాకే స్పందిస్తానని అన్నారు. విజయవాడలోని రమేశ్ ఆసుపత్రికి వెళ్లి అక్కడ చికిత్స పొందుతున్న ఎమ్మెల్సీ బచ్చుల అర్జునుడుని చంద్రబాబు పరామర్శించి అనంతరం మీడియాతో మాట్లాడుతూ .. బచావత్ ట్రైబ్యునల్కు, గెజిట్కు ఉన్న వ్యత్యాసాలను గుర్తించాల్సి ఉందని ఆయన అన్నారు. అయితే, ఈ విషయాలను ప్రస్తావించకుండా వైస్సార్సీపీ ప్రభుత్వం తప్పించుకునే ప్రయత్నం చేస్తోందని వివమర్శించారు. ఏపీ పట్ల సీఎం జగన్ బాధ్యత లేకుండా వ్యవహరిస్తున్నారని, తాము మాత్రం ఏపీ ప్రయోజనాల కోసం పోరాడతూనే ఉంటామని ఆయన చెప్పుకొచ్చారు.
Translation	Amaravati: In the wake of the water dispute between the two Telugu states, the Union Jal Shakti Ministry has released a gazette notification finalising the limits of the Krishna and Godavari river water boards. On this, the TDP chief Chandrababu Naidu responded. He said he would respond only after a thorough study of the gazette. Chandrababu went to the Ramesh Hospital in Vijayawada and visited MLC Bachula Arjunudu, who is undergoing treatment there, and later spoke to the media. He said the differences between the Bachawat Tribunal and the Gazette need to be identified. However, he said that the YSRCP government was trying to avoid mentioning these issues. He said that CM Jagan is acting irresponsibly towards AP and they will continue to fight for the interests of AP.
Headline	ఏపీ ప్రయోజనాల కోసం పోరాడతూనే ఉంటాం
Translation	We will continue to fight for the interests of AP
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 4.2: 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 స్క్రీన్ప్లే అందించిన "భీమ్లా నాయక్" సినిమాలో సంయుక్త మీనన్ రానా భార్య పాత్రలో కనిపించనుంది. ఈ సినిమాలో తన నటనకు ఫిదా అయిన త్రివిక్రమ్ ఆమెను మహేష్ బాబు సినిమాలో కూడా ఎంపిక చేసినట్లు తెలుస్తోంది.
Translation	Director Trivikram's habit is to put an elder sister or sister to the heroine whether it is necessary or not. In a way, this is one of the sentiments that Trivikram follows. Trivikram used the same sentiment in films like Jalsa, Attarintiki Daredi and Aravinda Sametha. Those films became blockbusters. However, according to the latest reports, Trivikram is going to use the same sentiment in his next film as well. It is known that Trivikram is going to do a film with Mahesh Babu in the lead role. Pooja Hegde is playing the female lead in the film. According to the latest reports, Samyuktha Menon will be seen as Pooja Hegde's sister in the film. Samyuktha Menon will be seen essaying the role of Rana's wife in "Bheemla Nayak", which is scripted by Trivikram. Apparently, Trivikram, who was impressed by her performance in the film, has also roped in her for Mahesh Babu's film.
Headline	మహేష్ బాబు సినిమాలో హీరోయిన్ గా రానా వైఫ్
Translation	Rana's wife as heroine in Mahesh Babu's film
Explanation	The article says Samyuktha Menon (who acted as Rana's wife in Bheemla Nayak movie) to act along with Mahesh Babu in a movie directed by Trivikram . However, the headline says Rana's wife as heroine in Mahesh Babu's movie which is misleading because it deviates from the core information present in the article.

Figure 4.3: Example of Least Related Headline

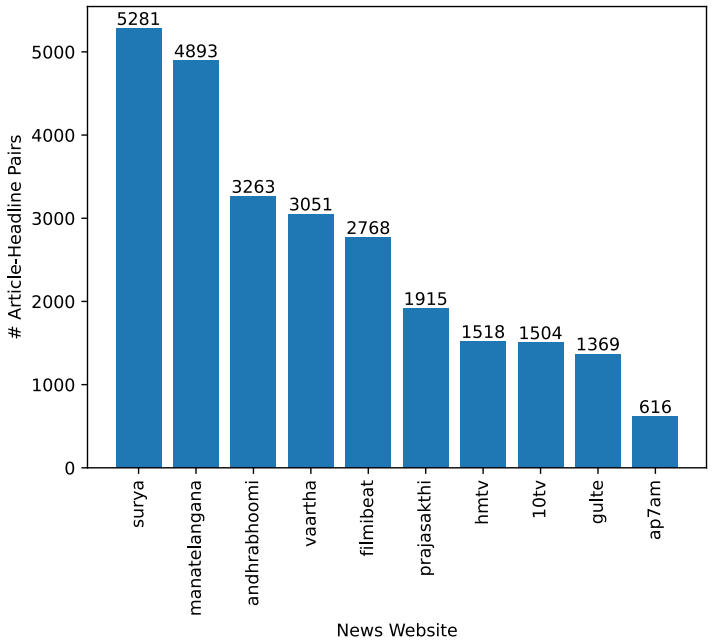

Figure 4.4: News website distribution in TeClass

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

Table 4.1: Category-wise counts in each data split

4.1.3 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 4.4, and Figure 4.5 respectively.

Data Splits: We allocated 70% for training, 15% for development and 15% for testing. To ensure unbiased performance and prevent category bias, we applied stratified sampling techniques. This ensures even distribution of articles from all 3 categories across the training, development, and test sets. The category-wise counts in each data split are presented in Table 4.1. Further statistical details of the TeClass dataset are available in Table 4.2.

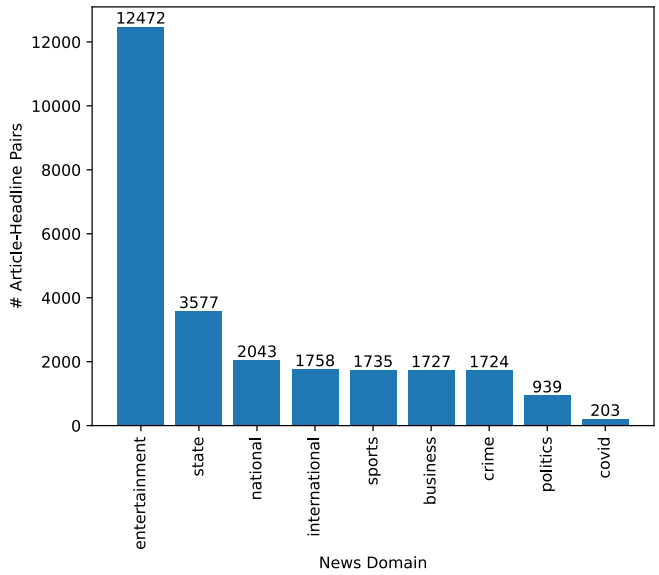

Figure 4.5: News domain distribution in TeClass

Inter-Annotator Agreement: Having multiple annotators (typically three or more) for annotation tasks is vital for several reasons. They enable the measurement of inter-annotator agreement, helping to identify and address ambiguous or challenging cases. Multiple annotators also help mitigate individual bias and promote a balanced, objective annotation process ensuring the robustness and quality of the annotated dataset. We use Fleiss' Kappa metric proposed by [26] and it resulted in an encouragingly high score of 0.77, indicating a substantial agreement among the annotators.

4.2 Experiments

We experimented with headline generation by using mT5 model trained on Mukhyansh Telugu dataset (refer section 3.2 in chapter 3). This was further fine-tuned on different subsets of TeClass to evaluate the impact of class-specific fine-tuning on the headline generation task. As seen in Table 4.3, non-fine-tuned model performs well enough but if we want the most relevant headline generation then class-aware training always significantly improves (5 points) ROUGE-L score across the board. In a human evaluation conducted by two volunteers on 50 news articles, we found that 34, 1, and 3 generated headlines were marked as FME, FSE, and STC respectively.

It is interesting to note that the best performance on all the relevant classes (FME, STC, FSE) is achieved by fine-tuning either on FME class or the combination of all the relevant classes. It is also interesting to see that the performance gain is not proportional to the training data size. In fact, we see a marked decrease in performance when all of the data is used. The best performance is achieved using 43% of the data (FME).

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

Table 4.2: TeClass Statistics

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

Chapter 5

Conclusion and Future work

5.1 Conclusion

Headline generation in low-resource languages, such as Indian languages, faces significant challenges due to the scarcity of large, high-quality datasets. Our work addresses this gap by introducing Mukhyansh, a comprehensive multilingual dataset scraped from the web with meticulous attention to each website's structure to avoid information loss. The Mukhyansh dataset comprises 3.39 million article-headline pairs covering eight Indian languages, making it approximately 13 times larger than the Indic language section of XL-Sum and five times larger than the IndicNLG-HG filtered dataset. The importance of our work is substantiated by empirical analysis of existing works, revealing issues such as data contamination, duplication, and other critical data quality issues. We establish strong baseline models for headline generation using the Mukhyansh dataset. Our approach involves training the models from scratch using a recurrent neural network-based encoder-decoder architecture. Furthermore, we fine-tune pre-trained multilingual models such as mT5-small and SSIB (IndicBARTSS) to improve their performance. Through extensive experimentation, we demonstrate the superiority of Mukhyansh and our SSIB baseline model, surpassing all existing works in Indian language headline generation.

We introduce "TeClass", a high-quality human-annotated dataset tailored for the task of relevance-based news headline classification in Telugu. The dataset comprises 26,178 article-headline pairs, meticulously annotated into three primary classes: Highly Related, Moderately Related, and Least Related. For the task of relevance-based headline generation, we further finetuned mT5 model (trained using Mukhyansh Telugu dataset) on different subsets of TeClass. Through experiments, we demonstrate the effectiveness of class-specific fine-tuning, highlighting the importance of relevance-based classification in improving the headline generation quality. Our results show that the best performance(increase in 5 ROUGE-L points) is achieved when fine-tuning on the class of most relevant headlines, even with a smaller subset of the data.

5.2 Ethics Statement

The distribution of the dataset collected from the web raises ethical considerations. We acknowledge that the copyright of the news articles collected from various websites remains with the original creators. Given that each website may have its own policies regarding data distribution or public availability, the released datasets will be licensed under CC BY-NC 4.0 DEED [1].

5.3 Limitations and Future Work

This section outlines the limitations of our current work and suggests avenues for future research to address these challenges and enhance headline generation in low-resource languages.

1. **Multilingual Model Fine-Tuning:** The limited availability of compute resources prevented us from fine-tuning any multilingual models. This limitation presents both a challenge and an opportunity for future research. We anticipate that fine-tuning multilingual models on the Mukhyansh dataset could lead to the development of new state-of-the-art models in headline generation.

2. **Extension to More Languages:** One primary limitation of Mukhyansh is its language coverage, currently including only 8 out of the 22 languages with official status in India. To overcome this limitation, we recommend expanding the dataset to encompass more Indian languages as well as languages from other low-resource regions.

3. **Fine-tuning for Specific Domains**: Fine-tuning headline generation models for specific domains or topics within Indian languages. This could involve creating specialized datasets or adapting existing models to better capture the nuances of different subject areas such as politics, sports, entertainment, finance or healthcare.

4. **Robust Evaluation Metrics:** Develop or adapt evaluation metrics that are better suited to assess the quality of headlines generated in low-resource languages. Traditional metrics like ROUGE may not capture all aspects of headline quality in these contexts, so exploring alternative metrics or refining existing ones could be valuable.

[1] https://creativecommons.org/licenses/by-nc/4.0/deed.en

Related Publications

- 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 (PACLIC-2023), pages 620–634, Hong Kong, China. Association for Computational Linguistics.
- Gopichand Kanumolu, Lokesh Madasu, Nirmal Surange, and Manish Shrivastava. **"TeClass: A Human-Annotated Relevance-based Headline Classification and Generation Dataset for Telugu**" LREC-COLING 2024.

Other Publications

- Gopichand Kanumolu, Lokesh Madasu, Pavan Baswani, Ananya Mukherjee and Manish Shrivastava "**Unsupervised Approach to Evaluate Sentence-Level Fluency: Do We Really Need Reference?**" IJCNLP-AACL SEALP-2023 Workshop.
- 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** " Findings of the Association for Computational Linguistics: ACL 2024.

Bibliography

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] M. Banko, V. O. Mittal, and M. J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325, Hong Kong, Oct. 2000. Association for Computational Linguistics.

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

[4] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. pages 103–111, Oct. 2014.

[5] S. Chopra, M. Auli, and A. 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, San Diego, California, June 2016. Association for Computational Linguistics.

[6] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. Khapra, and P. Kumar. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] B. Dorr, D. Zajic, and R. Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. Technical report, MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2003.

[9] D. K. Evans, J. L. Klavans, and K. McKeown. Columbia newsblaster: Multilingual news summarization on the web. In *Demonstration Papers at HLT-NAACL 2004*, pages 1–4, 2004.

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

[11] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[12] J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[13] X. Gu, Y. Mao, J. Han, J. Liu, H. Yu, Y. Wu, C. Yu, D. Finnie, J. Zhai, and N. Zukoski. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*, 2020.

[14] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. 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, Aug. 2021. Association for Computational Linguistics.

[15] B. Heinzerling and M. Strube. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] D. Jin, Z. Jin, J. T. Zhou, L. Orii, and P. 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*, pages 5082–5093, Online, July 2020. Association for Computational Linguistics.

[18] A. Kumar, H. Shrotriya, P. Sahu, A. Mishra, R. Dabre, R. Puduppully, A. Kunchukuttan, M. M. Khapra, and P. Kumar. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational*

Natural Language Learning, pages 280–290, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[23] C. Napoles, M. Gormley, and B. Van Durme. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada, June 2012. Association for Computational Linguistics.

[24] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

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

[27] A. M. Rush, S. Chopra, and J. 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, Sept. 2015. Association for Computational Linguistics.

[28] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. MLSUM: The multilingual summarization corpus. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, Nov. 2020. Association for Computational Linguistics.

[29] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[30] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. 27, 2014.

[31] A. Urlana, N. Surange, P. Baswani, P. Ravva, and M. Shrivastava. TeSum: Human-generated abstractive summarization corpus for Telugu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5712–5722, Marseille, France, June 2022. European Language Resources Association.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] M. Völske, M. Potthast, S. Syed, and B. Stein. TL;DR: Mining Reddit to learn automatic summarization. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. 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, Online, Oct. 2020. Association for Computational Linguistics.

[35] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[36] J. Zhang, Y. Zhao, M. Saleh, and P. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR, 2020.