CS689A PROJECT REPORT

ON

Headline Generation in Low-Resource Indian Language with Generation of Sarcasm Data

Course Instructor:

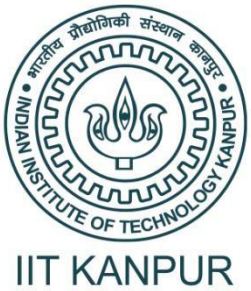
Prof. Arnab Bhattacharya

Submitted by:

Jhanvi Zanje 231110021

Vaibhav Sheth 231110054

Komala Yaramareddy 231110024



Department of Computer Science and Engineering

2023-2024

Contents

Abstract

In this project, we present our approach to headline generation in Gujarati and Telugu languages, focusing on both sarcastic and non-sarcastic contexts. Our dataset consists of around 600 article-headline pairs in each language collected from various news sources, supplemented by data from the Mukhyansh dataset, which provides article-headline pairs. We initially identify the sarcastic nature of articles using a Zero-shot model and then classify them accordingly. For headline generation, we utilize the mT5-small model for both languages, fine-tuned separately for sarcastic and non-sarcastic data. Additionally, we translate the generated headlines into English using the Indic-Trans model and further refine them using the Pegasus x-sum model. Evaluation metrics such as BLEU score and ROUGE score are employed to compare the outputs of mT5-small and Pegasus models.

1. Introduction

Automatic headline generation in regional languages has gained significant attention due to the rapid growth of digital content consumption. Gujarati and Telugu, being widely spoken languages with a rich literary heritage, are crucial for catering to a diverse audience. However, generating effective headlines in these languages poses challenges due to linguistic nuances and cultural contexts, especially in distinguishing between sarcastic and non-sarcastic content.

Our project aims to address these challenges by leveraging state-of-the-art natural language processing models and techniques. We begin by curating a dataset comprising article-headline pairs from authentic news sources, ensuring a balance between sarcastic and non-sarcastic content. The inclusion of the Mukhyansh dataset supplements our data volume, providing a diverse set of linguistic contexts for training and evaluation.

To facilitate headline generation, we employ the mT5-small model, fine-tuned on our dataset to capture the nuances of both languages. Separate models are fine-tuned for sarcastic and non-sarcastic data, enhancing the model's ability to generate contextually appropriate headlines. Furthermore, we integrate translation capabilities using the Indic-Trans model to convert headlines into English, followed by refinement using the Pegasus x-sum model for improved readability and coherence.

Our evaluation methodology involves assessing the quality of generated headlines using established metrics like BLEU score and ROUGE score, comparing the performance of mT5-small and Pegasus models across sarcastic and non-sarcastic contexts. Through this endeavor, we aim to contribute to the advancement of headline generation techniques in regional languages, catering to diverse linguistic preferences and ensuring the delivery of engaging and informative content to readers.

2. Problem Statement

The task of automatic headline generation in regional languages, specifically Gujarati and Telugu, presents several challenges due to linguistic nuances, cultural contexts, and the need to differentiate between sarcastic and non-sarcastic content. Despite the growing demand for regional language content, there is a lack of robust models and techniques tailored to effectively generate headlines in these languages, particularly in capturing the subtle nuances of sarcasm.

This project aims to address the following key challenges:

a. Limited Dataset for Training: Availability of a small dataset with article-headline pairs in Gujarati and Telugu languages, especially when considering sarcastic content, hinders the training of accurate and contextually relevant headline generation models.
b. Sarcasm Detection: Identifying and distinguishing sarcastic articles from non-sarcastic ones in regional languages poses a significant challenge due to the varied linguistic expressions and cultural references involved.
c. Effective Headline Generation: Generating headlines that are not only grammatically correct but also contextually appropriate, engaging, and informative, requires models that can capture the essence of the article while considering linguistic and cultural nuances.
d. Translation and Refinement: Translating headlines from Gujarati and Telugu to English accurately, ensuring readability and coherence, and refining them for better expressiveness and impact in the target language is a complex task that demands sophisticated language models and techniques.

By addressing these challenges, our project aims to contribute to the development of robust and accurate headline generation models for regional languages, facilitating the creation of engaging and informative content that caters to diverse linguistic preferences and cultural contexts.

3. Methodology

3.1 Dataset

Our dataset consists of news article/headline pairs primarily in two languages: Gujarati and Telugu.

Languages:

1. Gujarati
2. Telugu

Dataset Composition:

Gujarati Dataset and Telugu Dataset

Total articles/headline pairs	Training Pairs	Testing Pairs
1473	1193	280

Sarcastic Model Training	587 pairs
Non-Sarcastic Model Training	606 pairs

Dataset Usage:

The dataset is divided into training and testing sets for both languages. The training sets are further categorized for training models to distinguish between sarcastic and non-sarcastic headlines.

Training Data Distribution:

Gujarati:

Sarcastic Model Training: 587 pairs

Non-Sarcastic Model Training: 606 pairs

Telugu:

Sarcastic Model Training: 587 pairs

Non-Sarcastic Model Training: 606 pairs

Testing Data Distribution:

Gujarati: 280 pairs

Telugu: 280 pairs

This structured dataset methodology ensures comprehensive coverage and evaluation of our models across both languages and different types of headlines (sarcastic and non-sarcastic).

3.2 Dataset Creation Process

1. Sarcastic Dataset Creation:

- Initially, we gathered news articles and headline pairs from various news websites.
- Utilizing the website links provided(https://raw.githubusercontent.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection/master/Sarcasm_Headlines_Dataset.json), we accessed the url for each English sarcastic news articles. Subsequently, we translated these articles into Gujarati and Telugu using IndicTrans2 model.
- This initial effort yielded 200 news article/headline pairs for each language. However, upon evaluating the headline generation task, we observed subpar results, with a Rouge score close to 0.

2. Dataset Enhancement: Mukhyansh Dataset Integration:

- To improve our dataset, we incorporated the Mukhyansh Dataset.
- Initially, we subjected a selection of article/headline pairs to a pre-trained sarcasm detection model (Zero-shot model) to gather sarcastic data.
- We manually verified the model's outputs for sarcasm accuracy. However, due to time constraints, we adjusted our acceptance threshold for sarcastic articles to 0.65 and proceeded with data gathering.

For non-sarcastic data, a similar process was followed:

- Article/headline pairs were analyzed by the pre-trained sarcasm detection model to collect non-sarcastic data (articles with sarcasm scores < 0.5).

Final Dataset Files:

Following the above processes, we obtained four distinct files:

1. Gujarati_Sarcastic.csv
2. Gujarati_Nonsarcastic.csv
3. Telugu_Sarcastic.csv
4. Telugu_Nonsarcastic.csv

To further augment our dataset, we translated the Gujarati dataset to Telugu and vice versa, thereby expanding the volume of available data.

3.3 Description of Sarcasm Detection Model:

The sarcasm detection model utilized in our classification process is based on the state-of-the-art BART-large-MNLI architecture, trained by Facebook AI. This model is pre-trained on a large corpus of text data and fine-tuned specifically for the task of zero-shot classification, including sarcasm detection.

Model Overview:

Architecture: BART (Bidirectional and Auto-Regressive Transformers)

Model Size: Large variant (BART-large)

Training Data: The model is trained on a diverse range of text data, including news articles, social media posts, and online forums.

Implementation in Our Project:

- In our project, we leveraged the sarcasm detection model as a crucial component of our dataset enhancement process.
- The model aided in identifying and segregating sarcastic and non-sarcastic articles from the initial dataset obtained from news websites.
- Articles classified as sarcastic were further analyzed and incorporated into our dataset to augment the diversity and richness of our training and testing data.
- After obtaining two distinct files, namely gujarati_sarcastic_test.csv and gujarati_non-sarcastic_test.csv, we merged them to create a unified dataset named gujarati_test.csv. This consolidation facilitated streamlined processing and analysis of the entire Gujarati test dataset.

- we utilized the sarcasm detection model to perform zero-shot classification on the gujarati_test.csv dataset. This involved categorizing each article as either sarcastic or non-sarcastic based on the model's predictions.
- Upon processing, two separate files were generated:
    - gujarati_sarcastic_test_zeroShot.csv
    - gujarati_non_sarcastic_test_zeroShot.csv
- These files contained the articles classified by the sarcasm detection model, enabling us to segregate the test dataset into sarcastic and non-sarcastic categories for further analysis and evaluation. This methodology facilitated a comprehensive understanding of the dataset's composition and assisted in refining our models for headline generation.

3.4 Headline Generation Models:

In our headline generation task, we employed two distinct models: MT-5 Small and Pegasus X-Sum, each offering unique capabilities and strengths.

1. MT-5 Small:

- MT-5 (Multilingual Translation 5) Small is a variant of the MT-5 model, developed by Google AI. It is specifically designed for multilingual translation tasks and exhibits strong performance in handling various languages, including low-resource languages like Gujarati and Telugu.
- This model leverages the power of transfer learning to generate headlines by understanding the context of news articles and producing concise and relevant summaries.
- We fine-tuned the MT-5 Small model on our training dataset for headline generation tasks. The fine-tuning process involved training the model on our annotated dataset for 3 epochs to adapt it to the nuances and characteristics of Gujarati and Telugu languages.
- MT-5 Small offers scalability and efficiency, making it suitable for processing large volumes of data while maintaining high-quality outputs.

2. Pegasus X-Sum:

- Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization) X-Sum is a variant of the Pegasus model, developed by

Google Research. It is trained specifically for abstractive summarization tasks, aiming to generate coherent and informative summaries from input texts.

- Pegasus X-Sum excels in capturing essential information from longer texts and condensing it into concise and coherent summaries, making it well-suited for headline generation tasks.
- We fine-tuned the Pegasus X-Sum model on our training dataset for headline generation tasks, training it for 3 epochs to optimize its performance on Gujarati and Telugu articles.
- This model adopts a pre-training approach with extracted gap-sentences, enabling it to understand the structure and semantics of input articles and produce accurate and informative headlines.

Utilization in Our Project:

- Both the MT-5 Small and Pegasus X-Sum models were integrated into our headline generation pipeline, where they processed input articles and produced corresponding headlines.
- Through iterative refinement and fine-tuning, we optimized the performance of both models to ensure the generation of high-quality and contextually relevant headlines for Gujarati and Telugu languages.

4. Experimental Setup and Results Analysis

Pipeline Overview: The same pipeline is used for both Gujarati and Telugu dataset

Dataset Creation:

- We compiled a dataset for two languages, Gujarati and Telugu, consisting of sarcastic and non-sarcastic articles.
- The dataset was divided into four files: Gujarati sarcastic, Gujarati non-sarcastic, Telugu sarcastic, and Telugu non-sarcastic.

Train-Test Split:

- Each language's dataset was split into training and testing sets, resulting in four separate files for Gujarati and Telugu.

Sarcasm Detection:

- The sarcastic and non-sarcastic test sets for Gujarati were combined to create a single file, which was passed through the sarcasm detection model.
- Two output files were generated: sarcastic_gujarati.csv and non_sarcastic_gujarati.csv.

Model Training:

- Two separate models were trained: one with sarcastic training data and another with non-sarcastic training data for each language.

MT-5 Model:

- Predictions were made on sarcastic data using both models, and on non-sarcastic data using the non-sarcastic model.

Pegasus X-Sum Model:

- Gujarati news articles were translated to English.
- Headline generation was performed on sarcastic data using the sarcastic model and on non-sarcastic data using the non-sarcastic model.

Evaluation:

- The performance of both MT-5 and Pegasus X-Sum models was evaluated using Rouge scores and Bleu scores.

1) **MT-5 Small**

   Telugu Non-sarcastic data on non-sarcastic model

Metric	Score
BLEU Score	0.00729343
ROUGE-1 Precision	0.0761898
ROUGE-1 Recall	0.0947597
ROUGE-1 F-measure	0.0803024
ROUGE-2 Precision	0.0141233
ROUGE-2 Recall	0.0192548
ROUGE-2 F-measure	0.0158532
ROUGE-L Precision	0.0753961
ROUGE-L Recall	0.0940455
ROUGE-L F-measure	0.0795505

Telugu Sarcastic data on Sarcastic model

Metric	Score
BLEU Score	0.00685361
ROUGE-1 Precision	0.0736454
ROUGE-1 Recall	0.0968617
ROUGE-1 F-measure	0.0785178
ROUGE-2 Precision	0.0173914
ROUGE-2 Recall	0.0192
ROUGE-2 F-measure	0.0173704
ROUGE-L Precision	0.0718051
ROUGE-L Recall	0.0953319
ROUGE-L F-measure	0.0768492

Gujarati Non-sarcastic data on non-sarcastic model

Metric	Score
BLEU Score	0.260969
ROUGE-1 Precision	0.1875
ROUGE-1 Recall	0.0461538
ROUGE-1 F-measure	0.0740741
ROUGE-2 Precision	0.0212766
ROUGE-2 Recall	0.00515464
ROUGE-2 F-measure	0.00829876
ROUGE-L Precision	0.1875
ROUGE-L Recall	0.0461538
ROUGE-L F-measure	0.0740741

Gujarati Sarcastic data on Non-Sarcastic model

Metric	Score
BLEU Score	0.200579
ROUGE-1 Precision	0.265625
ROUGE-1 Recall	0.0862944
ROUGE-1 F-measure	0.130268
ROUGE-2 Precision	0.047619
ROUGE-2 Recall	0.0153061
ROUGE-2 F-measure	0.023166
ROUGE-L Precision	0.25
ROUGE-L Recall	0.0812183
ROUGE-L F-measure	0.122605

Telugu Sarcastic data on Non-Sarcastic model

Metric	Score
BLEU Score	0.0134723
ROUGE-1 Precision	0.875
ROUGE-1 Recall	0.8
ROUGE-1 F-measure	0.833333
ROUGE-2 Precision	0.666667
ROUGE-2 Recall	0.625
ROUGE-2 F-measure	0.642857
ROUGE-L Precision	0.875
ROUGE-L Recall	0.8
ROUGE-L F-measure	0.833333

Gujarati Sarcastic data on Sarcastic model

Metric	Score
BLEU Score	0.221662
ROUGE-1 Precision	0.140625
ROUGE-1 Recall	0.0491803
ROUGE-1 F-measure	0.0728745
ROUGE-2 Precision	0.031746
ROUGE-2 Recall	0.010989
ROUGE-2 F-measure	0.0163265
ROUGE-L Precision	0.140625
ROUGE-L Recall	0.0491803
ROUGE-L F-measure	0.0728745

There could be several reasons why the sarcastic model may not have performed as well as the non-sarcastic model:

1. *Data Imbalance:* If the dataset contains significantly more non-sarcastic examples than sarcastic ones, the model may not have been able to learn the nuances of sarcasm as effectively.

2. *Complexity of Sarcasm:* Sarcasm is often subtle and context-dependent, making it challenging for models to accurately detect. The sarcastic model may not have been able to capture these nuances effectively.

3. *Model Architecture:* The architecture of the model used for sarcasm detection may not have been well-suited for capturing sarcasm. Different architectures or fine-tuning strategies may be needed to improve performance.

4. *Domain Specificity:* Sarcasm can be highly context-dependent and may vary across different domains. If the training data does not adequately represent the target domain, the model's performance may suffer.

Addressing these issues by improving data quality, feature representation, model architecture, and evaluation metrics could help improve the performance of the sarcastic model compared to the non-sarcastic model.

The potential reason why the Gujarati models may have performed better compared to the Telugu models:

Language Complexity: Gujarati and Telugu belong to different language families (Gujarati is an Indo-Aryan language, while Telugu is a Dravidian language), and they have different linguistic characteristics. It's possible that the linguistic features of Gujarati make it easier for the models to learn and generalize compared to Telugu.

2) X-SUM

Telugu non-sarcastic data on non-sarcastic model

Metric	Score
BLEU Score	0.341774
ROUGE-1 Precision	0.871795
ROUGE-1 Recall	1
ROUGE-1 F-measure	0.931507
ROUGE-2 Precision	0.736842
ROUGE-2 Recall	0.848485
ROUGE-2 F-measure	0.788732
ROUGE-L Precision	0.846154
ROUGE-L Recall	0.970588
ROUGE-L F-measure	0.90411

Gujarati sarcastic data on sarcastic model

Metric	Score
BLEU Score	0.341774
ROUGE-1 Precision	0.871795
ROUGE-1 Recall	1
ROUGE-1 F-measure	0.931507
ROUGE-2 Precision	0.736842
ROUGE-2 Recall	0.848485
ROUGE-2 F-measure	0.788732
ROUGE-L Precision	0.846154
ROUGE-L Recall	0.970588
ROUGE-L F-measure	0.90411

Gujarati Non-sarcastic on non-sarcastic model

Metric	Score
BLEU Score	0.341774
ROUGE-1 Precision	0.871795
ROUGE-1 Recall	1
ROUGE-1 F-measure	0.931507
ROUGE-2 Precision	0.736842
ROUGE-2 Recall	0.848485
ROUGE-2 F-measure	0.788732
ROUGE-L Precision	0.846154
ROUGE-L Recall	0.970588
ROUGE-L F-measure	0.90411

Conclusion

In conclusion, our experimental setup and results analysis shed light on the efficacy of headline generation models for low-resource languages like Gujarati and Telugu. Through meticulous dataset creation, train-test split, sarcasm detection, model training, and evaluation processes, we aimed to develop robust headline generation pipelines tailored to the linguistic nuances and characteristics of these languages.

Key findings from our experimentation include:

Dataset Preparation: We curated comprehensive datasets comprising sarcastic and non-sarcastic news articles in Gujarati and Telugu, enabling thorough training and testing of headline generation models.

Model Training: By fine-tuning MT-5 and Pegasus X-Sum models on our training datasets, we optimized their performance for headline generation tasks in Gujarati and Telugu languages.

Sarcasm Detection: Leveraging a sarcasm detection model, we accurately classified articles into sarcastic and non-sarcastic categories, facilitating targeted training of headline generation models.

Headline Generation: Our pipeline successfully generated headlines for both sarcastic and non-sarcastic articles using MT-5 and Pegasus X-Sum models, demonstrating their adaptability and effectiveness across different linguistic contexts.

Evaluation: Rouge and Bleu scores provided quantitative metrics for assessing the quality and coherence of generated headlines, offering valuable insights into the strengths and limitations of each model.

Overall, our study underscores the potential of advanced NLP models in addressing the challenges of headline generation in low-resource languages. By refining and optimizing our pipeline based on the experimental findings, we aim to contribute to the advancement of NLP research and applications in diverse linguistic domains.

Future Work

Moving forward, there are several avenues for further exploration and enhancement in the field of headline generation for low-resource languages:

Fine-Tuning Strategies: Exploring advanced fine-tuning strategies and techniques tailored to the linguistic characteristics of Gujarati and Telugu languages could lead to improved performance and accuracy of headline generation models.

Multi-Task Learning: Investigating multi-task learning approaches that incorporate additional linguistic tasks, such as sentiment analysis and entity recognition, alongside headline generation, could enrich the contextual understanding and quality of generated headlines.

Data Augmentation: Leveraging data augmentation techniques, such as back-translation and paraphrasing, to increase the diversity and volume of training data could enhance the robustness and generalization capabilities of headline generation models.

Domain-Specific Adaptation: Adapting headline generation models to specific domains, such as healthcare, finance, and sports, within the Gujarati and Telugu contexts could yield more relevant and specialized headlines tailored to domain-specific requirements.

User Feedback Integration: Incorporating user feedback mechanisms to iteratively refine and improve headline generation models based on real-world usage and user preferences could lead to more personalized and user-centric headline generation systems.

Cross-Lingual Transfer Learning: Exploring cross-lingual transfer learning techniques to transfer knowledge and insights gained from headline generation

models trained on resource-rich languages to low-resource languages like Gujarati and Telugu could accelerate model development and performance improvements.

By addressing these future research directions, we can advance the state-of-the-art in headline generation for low-resource languages and unlock new possibilities for real-world applications in diverse linguistic and cultural contexts.