Classification and Correction of Non-Representative News Headlines

Gail Batutis
Center for Data Science
New York University
gsb8013@nyu.edu

Annabelle Huether
Center for Data Science
New York University
amh9750@nyu.edu

Mary Nwangwu
Center for Data Science
New York University
mcn8851@nyu.edu

Allison Redfern
Center for Data Science
New York University
amr10211@nyu.edu

Mallory Sico
Center for Data Science
New York University
mcs9834@nyu.edu

Abstract

This project addresses the issue of inaccurate headlines in news media by developing a model to classify headline-article alignment and suggest improvements for misaligned cases. The focus is on combating misinformation, particularly in topics like politics, crime, and clickbait articles. The study employs a nuanced approach to model construction and scoring, emphasizing the F1 score to handle class imbalance. A generative model, specifically fine-tuning a pretrained model for headline generation, is introduced, marking a notable advancement. The headline classification model demonstrates robust performance, with BERT fine-tuned on non-summarized text consistently outperforming in various scenarios. Evaluation of the headline generation models favors T5 for its well-rounded performance in relevance, readability, and style. GitHub Repository.

1 Introduction

This project aims to address a significant problem in news media: the frequent inaccuracy of headlines in representing article content. Our primary goal is to develop a model that classifies whether headlines accurately match their respective articles. For those that do not align, the model will suggest improved headlines. This objective is crucial for combating the spread of misinformation through headlines, especially in influential areas like politics, crime, and clickbait articles. By empowering news consumers to discern the accuracy of headlines, we aim to promote informed decision-making in a media landscape where headlines strongly influence public perception. The focus is on countering the dissemination of fabricated or sensationalized headlines, fostering a more factual and informed approach to understanding current issues.

Previous explorations of this problem in related work have attempted to classify headlines with a highly unbalanced dataset and an ungeneralizable scoring metric. In this study, a similar challenge is tackled but with a nuanced approach to address issues related to model construction and scoring, particularly emphasizing the use of the F1 score to effectively handle class imbalance. Furthermore, we've expanded the scope by integrating a generative model, a novel addition to previous work with this dataset. This extension involves fine-tuning a pretrained model for headline generation, marking a notable step forward in our approach.

The headline classification modeling task demonstrates robust performance across various scenarios. A gradient boosted decision trees baseline using Word2Vec embeddings on the Fake News Challenge Dataset with full-text articles provides a solid starting point. Furthermore, fine-tuned BERT (base-cased) models consistently outperform the baseline in headline classification across different dataset configurations (full text, long summaries, and short summaries). The headline generation model, evaluated through a sample human assessment, indicates notable performance. The T5 (base-en-generate-headline) model stands out for relevance, readability, and style, while the Llama-2 7B Chat model excels in the latter two aspects. Overall, the comprehensive evaluation favors T5 as the chosen model for its well-rounded performance in generating new headlines.

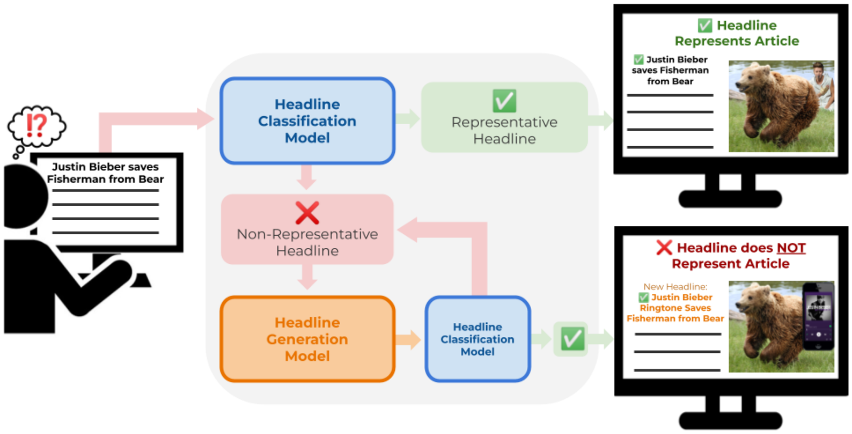

⁉
Justin Bieber saves Fisherman from Bear
Headline Classification Model
✅
Representative Headline
✅ Headline Represents Article
✅ Justin Bieber saves Fisherman from Bear
❌
Non-Representative Headline
Headline Generation Model
Headline Classification Model
✅
❌ Headline does NOT Represent Article
New Headline:
✅ Justin Bieber Ringtone Saves Fisherman from Bear

Figure 1: Objective flowchart displaying headline article moving through headline classification, then non-representative headlines flowing through headline generation and further classification.

2 Related Work

Preventing the spread of misinformation, specifically in headlines, has been a studied topic in recent years. The Fake News Challenge[1] in 2017 used Stance Detection to compare article bodies and headlines to determine relative positions in four classes: agrees, disagrees, discusses, and unrelated. The top model, a mix of gradient-boosted decision trees and a deep convolutional neural network, achieved a weighted score of 82.02. Notably, it focused on headline stance rather than representativeness. This project differs by addressing the accuracy of headlines in representing articles using human evaluation, and includes headline generation, not just classification. A 2021 paper[2] by Sepúlveda-Torres et al. tackles the headline/body text dissonance issue using the HeadlineStanceChecker, achieving 94.31 percent accuracy across agree, disagree, discuss, and unrelated categories. However, due to an unbalanced dataset, their results on agree and disagree categories were less favorable.

3 Approach

Figure 1 displays an overview of the objective of this project. A headline classification model and headline generation model were trained and tuned to allow for classification and correction of non-representative headlines for a set of articles. When combining these two model types as a system, a headline and article would first flow through the headline classification model (in blue) where the headline would be classified as representative of the article body or non-representative. If the headline is labeled as non-representative, the article body would flow through the headline generation model (in orange), where a new headline would be generated. To assess this new headline, the generated headline and article body would once again flow through the original headline classification model, and continue repeating this cycle until a representative headline is produced. To achieve this model system, we first establish a baseline for representativeness detection using gradient boosted decision trees with Word2Vec embeddings. We then adopt a two-stage network, consistent with previous work, where the first stage involves summarizing the article body, and the second stage classifies the headline based on this summarization using BERT, and evaluating model performance using F1 score. In the generation task, we focus on generating agreeable or discuss-worthy headlines based on article content, leveraging pretrained generation models, T5 and Llama-2. For evaluation, the generated headlines are treated as new data points, and we assess their quality using our classification model and human evaluation on relevance, creating a comprehensive approach to address both classification and generation aspects of our research on headline-article relationships.

4 Experiments

4.1 Data

The dataset used was sourced from the Emergent Dataset by Craig Silverman and was accessible on the FNC-1 GitHub[3]. Instances were structured as (headline, body, stance), with the stance representing categories: unrelated, discuss, agree, or disagree. Both the training and testing sets were organized in four separate CSV files, with one file each for ‘stances’ and ‘bodies’ in both sets. The combined dataset was 11.7 MB. To address class imbalance, ‘agree’ and ‘discuss’ were aggregated into one category, and ‘disagree’ and ‘unrelated’ into another, transforming the problem into a binary classification task. Additionally, text summarization techniques were employed to create concise representations of the article content. These preprocessing steps aimed to maximize the quality of the results in both the classification and generation tasks.

4.2 Evaluation method

To assess the model’s accuracy in classifying training headlines, F1 score was used. The F1 score, considering precision and recall, proved valuable due to the dataset’s class imbalance and the higher cost associated with false negatives. Unlike prior models discussed in the Related Work section, our model’s objective was binary classification, making a direct F1 score comparison impractical. Instead, the model’s F1 score was compared to the baseline model (gradient boosted trees). Achieving a high F1 score allowed the model to evaluate generated headlines for the negative class by determining their representativeness through model assessment.

After classifying the headlines, the subsequent step involved replacing non-representative headlines with newly generated ones. Two models, T5-Base Headline Generator and Llama-2 7B Chat, were tested. Extracting a subset of 50 headlines from each model, human evaluation was conducted. The headline evaluation rubric, as shown in Appendix A, assesses headlines based on four criteria: relevance, ensuring alignment with article content; discusses, gauging discussion of at least one of the subjects of the article; readability, examining clarity and grammatical correctness; and style, evaluating engagement and headline-like qualities. The selected model, determined by higher rates of relevance and discussion, facilitated the final step: evaluating generated headlines for the negative class by running them through the model to assess their representativeness. The other two metrics were chosen for marks of quality of headline generation and associated comparisons, but are not factors in the classification model.

4.3 Experimental details

A multifaceted approach was taken to the headline-body representativeness classification task. Consistent with prior literature, summarization of the article bodies in the dataset was performed. This was done using two different models, PEGASUS[4] and T5[5]. PEGASUS was pretrained to predict masked sentences allowing it to transfer well to summarization. The T5 model uses a text-to-text framework that applies well to summarization with simple prompt engineering. Long and short summaries were created with each model having 200 and 100 token length respectively. The models implemented maximum token limits of 512 and 1024 for PEGASUS and T5 respectively, meaning many of the articles were truncated before summarizing.

BERT has a next sentence prediction objective during pre-training that makes it useful for downstream tasks such as sequence to sequence classification[6]. Thus, BERT was chosen as the optimal model to fine-tune to perform headline-body representativeness classification in the news headline correction pipeline. Following summarization, the test set from the fake news challenge was split into validation and test sets, each containing unique article body IDs. A gradient boosted trees model was trained and hyperparameter tuned as the baseline model. Then, a BERT (base-cased)[7] model was fine-tuned with the same hyperparameters (3 epochs, with a batch size of 16, a learning rate of 2.00E-5, and a weight decay of 0.01) on the non-summarized article bodies training set, and also on each of the summarized article bodies training sets from the four summarization methods (all with the same headlines) to compare performance. During the fine-tuning of each model, BERT was evaluated after each epoch on F1 score on the non-summarized article bodies in the evaluation set.

The fine-tuned T5-Base[8] model utilized to generate headlines was trained on a collection of 500k articles with headings for the purpose of creating a one-line heading suitable for a given article. The experimental setup involved configuring beam outputs with the parameters: max length 64, 3 beams, and early stopping, ensuring an effective approach to headline generation. To fine-tune the Llama-2 7B Chat[9] model for headline generation, adaptations were made to accommodate low resource constraints. To address these constraints, a sharded version[10] of the 7B model was loaded and int4 quantization was applied. Implementing parameter-efficient fine-tuning (PEFT) with low-rank adaptors (LoRA), the model was trained with the Hugging Face autotrain framework[11] in Google Colab, utilizing its T4 GPU. The training process incorporated the following hyperparameters: learning rate 2.00E-4, 5 epochs, batch size 1, block size 1024, SFT trainer, warmup ratio 0.1, weight decay 0.01, gradient accumulation 4, LoRA r 16, LoRA alpha 32, and LoRA dropout 0.05. The model was then fine-tuned for generating headlines using the News Headlines dataset[12] from Hugging Face, which consists of news articles and their corresponding headlines.

4.4 Results

For headline classification, as shown in Appendix B, BERT fine-tuned on the non-summarized article bodies outperformed the baseline model and BERT models fine-tuned on summarized article bodies, with the shorter PEGASUS generated summarized bodies coming in close second. After some fine tuning of BERT on the non-summarized article bodies and corresponding headlines, the best BERT model (5 epochs, batch size 16, learning rate 2.00E-5, and weight decay 0.1) attained a final epoch validation F1 score of 0.9276 and a test F1 score of 0.9247 on the classification task. As shown in Appendix C, the highest performing BERT model achieved an area under the receiver operating characteristic curve (AUC) of 0.99 and reasonable calibration.

The results of the human evaluation of the newly generated headlines are presented in Table 1. The T5 model was more effective at creating headlines which discussed the topics of the article in general and which captured the main idea of the text. The Llama-2 model was better suited for creating readable and headline-style text. When the generated headlines from the T5 model were fed back into our classification model, 97.46% of them were classified as representative.

Table 1: Headline generation model human evaluation results

Model	**Relevant**	**Discusses**	**Readability**	**Style**
T5	**62%**	**100%**	72%	54%
Llama-2 7B Chat	58%	80%	**78%**	**80%**

After finding the best headline classification and generation model, these models were used as described in Section 3 and displayed in Figure 1 and assessed on the test data. From the test data, 12,806 articles and headlines were fed through the BERT classification model. Of those articles, 2,979 (23%) were classified as representative immediately, while 9,827 (77%) were classified as non-representative. Those articles with headlines classified as non-representative were fed through the T5 headline generation model and assigned new generated headlines. The 9,827 generated headlines and articles were then classified again by BERT. After this first pass through headline generation, 9,577 (97%) headlines were now classified as representative while only 250 (3%) were not.

5 Analysis

These results necessitate an evaluation of the methodologies employed in this news headline classification and correction study. The unexpected finding that summarizartion did not improve BERT's performance on the classification task could possibly be due to token-limit truncation removing relevant information in longer articles. Prompt engineering could be further explored to improve the models' summarization performance on this dataset specifically despite their pretraining on similar tasks. In addition, in this study, BERT was fine-tuned to classify examples in which the article headline agreed with or discussed what was in its corresponding body as the positive class. In future work, the focus of classifier fine-tuning could shift to an objective that incorporates more thematic relevance by possibly creating a new dataset or new labels to better capture this. In the analysis of our headline generation models, the human evaluation revealed distinct strengths and weaknesses in the headline generation capabilities of the T5 and Llama-2 models. Human evaluation indicated that the T5 model exhibited superior performance in terms of relevance to and discussion about the article body, but could sometimes generate dry summaries instead of headline-style text. On the other hand, the Llama-2 model excelled in overall readability and style, and crafted well-written and stylistically appropriate headlines. However, the Llama-2 model was more prone to generating headlines that occasionally lacked any relevance to the article content, or hallucinated specific parts. We noticed that Llama-2 headlines almost always started with a number, such as a year, age, or list ("Five Reasons Why. . . ") even if there was no number in the article body.

6 Conclusion

In conclusion, this research addresses the issue of non-representative headlines in news media. Our model, using BERT for headline classification and T5 for headline generation, demonstrates robust performance with a best F1 score of 0.9276. Human evaluation highlights the effectiveness of T5 in capturing the article's main idea. The application of these models yielded promising results, transforming non-representative headlines into representative ones. The potential impact on the news industry is substantial, as these models could serve as pre-publishing tools for news companies or as browser extensions, enhancing headline-article alignment and mitigating misinformation.

References

[1] Fake News Challenge. "Fake News Challenge Stage 1 (FNC-I): Stance Detection." Fake News Challenge, www.fakenewschallenge.org/.

[2] Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, Manuel Palomar. "HeadlineStanceChecker: Exploiting summarization to detect headline disinformation." Journal of Web Semantics, Volume 71, 2021, 100660. https://doi.org/10.1016/j.websem.2021.100660.

[3] FNC-1 GitHub Repository. https://github.com/FakeNewsChallenge/fnc-1.

[4] Jingqing Zhang, Yao Zhao, Mohammad Saleh, & Peter J. Liu. (2019). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.

[5] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.

[6] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. http://arxiv.org/abs/1810.04805.

[7] bert-base-cased. https://huggingface.co/bert-base-cased.

[8] Michau/t5-base-en-generate-headline. Hugging Face. https://huggingface.co/Michau/t5-base-en-generate-headline.

[9] meta-llama/Llama-2-7b-chat-hf. Hugging Face. https://huggingface.co/meta-llama/Llama-2-7b-chat-hf.

[10] TinyPixel/Llama-2-7B-bf16-sharded. Hugging Face. https://huggingface.co/TinyPixel/Llama-2-7B-bf16-sharded.

[11] Huggingface Autotrain-Advanced. https://github.com/huggingface/autotrain-advanced.

[12] valurank/News_headlines. Hugging Face. https://huggingface.co/datasets/valurank/News_headlines.

[13] Tunstall, L., L. v. Werra, and T. Wolf (2022). Natural language processing with transformers: building language applications with Hugging Face (First edition ed.). Sebastopol, CA: O'Reilly Media. OCLC: on1266359932.

[14] He, H., L. Jiang, W. Yuan, X. Pan, Y. Kuang, and D. Rothermel (2023). NYU DS-GA 1011 Calendar and Course Content. https://nyu-cs2590.github.io/fall2023/calendar/.

[15] Awan, A. A. (2023). Fine-Tuning LLaMA 2: A Step-by-Step Guide to Customizing the Large Language Model. https://www.datacamp.com/tutorial/fine-tuning-llama-2

7 Student Contributions

Gail Batutis, Annabelle Huether, Mary Nwangwu, Allison Redfern, and Mallory Sico: Project ideation, human evaluation, presentation, report
Gail Batutis: T5 generation model
Annabelle Huether: BERT finetuning
Mary Nwangwu: Llama-2 finetuning
Allison Redfern: Baseline model, visualizations
Mallory Sico: Summarization models

8 Appendix

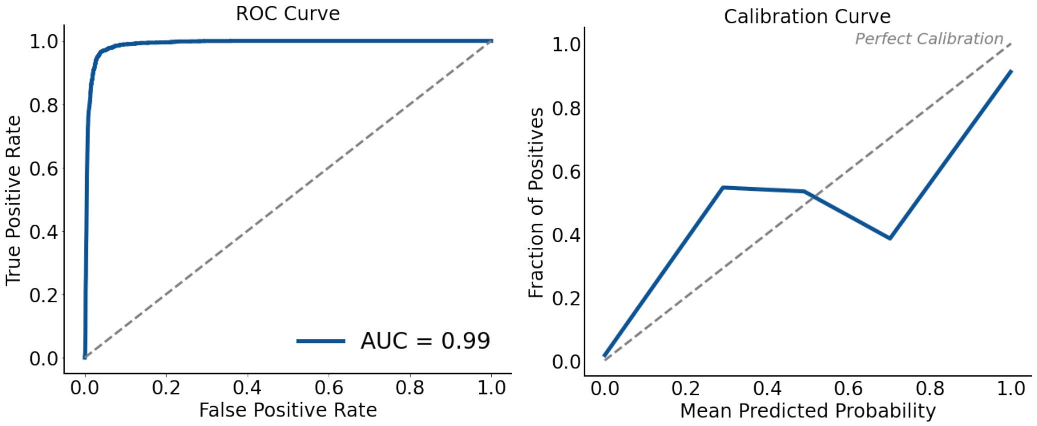
Appendix A. Headline Human Evaluation Rubric

Relevant	**Discusses**	**Readability**	**Style**
Relevant (1): The headline accurately reflects the main content of the article, providing a clear and meaningful representation of the key themes and topics.	Discusses (1): The headline discusses at least one of the subjects of the article	Readable (1): The headline is clear, easy to understand, and free from significant grammar or structure issues.	Headline Style (1): The headline exhibits a style reminiscent of news articles, with a good balance of clarity and engaging language, making it compelling for the target audience.
Not Relevant (0): The headline is not directly related to the article content or fails to capture the main themes and topics, resulting in a disconnect between the headline and the actual content.	Does not Discuss (0): The headline does not mention any of the topics in the article	Not Readable (0): The headline has readability issues, such as unclear phrasing, grammar errors, or structural mistakes.	Not Headline Style (0): The headline lacks a news-style flair, creativity, or engagement, reading more like a bland summary

Appendix B. Model Test F1-Score Comparison

Model	**Training Data**	**Test F1 Score**
Gradient Boosted Decision Trees (Baseline)	FNC Dataset - Full Text	0.8802
BERT (base-cased)	**FNC Dataset - Full Text**	**0.9245**
BERT (base-cased)	FNC Dataset - Pegasus Short Summaries	0.9219
BERT (base-cased)	FNC Dataset - Pegasus Long Summaries	0.9069
BERT (base-cased)	FNC Dataset - T5 Short Summaries	0.9052
BERT (base-cased)	FNC Dataset - T5 Long Summaries	0.9000

Appendix C. BERT Headline Classification Model ROC and Calibration Curves


ROC Curve
True Positive Rate
False Positive Rate
AUC = 0.99
Calibration Curve
Perfect Calibration
Fraction of Positives
Mean Predicted Probability