Midterm Report: Automatically Generating Corrective Context for Misleading News Headlines

Robert Wienröder
Grad Project in INFO 259 (Prof. David Bamman)
University of California, Berkeley
robertwienroeder@berkeley.edu

Abstract

News headlines have been playing a crucial role in shaping public opinion for many years now. But many headlines, especially nowadays, are intentionally sensationalist or clickbait and often leave out important context from the full article, which can influence people's perception on the given topic. Several research groups have been working on identifying such misleading headlines, but without further explanation for the reader. This midterm report presents the current state of my term project: A novel NLP system that not only detects misleading headlines, but also automatically generates clarifying context by extracting and summarizing relevant information from the article body. By bridging the gap between clickbait headlines and nuanced article content, this project work aims to improve media literacy and reduce misinformation.

1 Introduction

Often, the only part of an article that many readers see is the headline. So when the headline is misleading, it can influence public opinion in a potentially harmful way. To address this issue, my project develops a system that not only detects misleading headlines but also generates clarifying context to it. My approach is mainly built on three components: a detection module that classifies headlines as misleading or reliable using features such as tf-idf and sentiment scores; and a corrective context generation module that fine-tunes a generative model (e.g., a pointer-generator network, T5, or BART) to produce concise summaries clarifying the omitted details. I have furthermore looked into the possibilities of distant supervision to extend train sets for the generation task (I will have to determine whether this could be beneficial). The pipeline lays the foundation for improving media literacy of the public by making sure that readers the true meaning behind sensational headlines. Given the tight schedule and the challenges of working solo, the immediate goal is to develop a streamlined prototype that validates this approach.

2 Related Work

I have reviewed roughly 25–30 academic papers to position my project within the research landscape on this topic. I want to highlight the following in context of this work (see References section for the full list of references):

2.1 Detection and Explanation:

Shen et al. (2023) introduce a framework that detects headline hallucinations and generates natural language explanations by leveraging a model that is pretrained with natural language inference and augmented by explanations about why they were flagged. Although this seems to be similar to my approach, my explanations will not be about the reason for a headline being flagged, but about what is wrong or misleading.

2.2 Contradiction Recognition:

Sepulveda Torres et al. (2023) focus on contradiction recognition in Spanish news to identify when headlines are not reflecting the article content correctly. I leverage similar contradiction signals to flag misleading headlines and then generate summaries that rectify the discrepancy.

2.3 Hierarchical Encoding:

Yoon et al. (2022) propose a deep hierarchical encoder that processes news articles at both the word and paragraph levels using an Independent Paragraph (IP) data augmentation strategy. These techniques could be helpful for me to capture fine-grained relationships between headlines and articles.

2.4 Clickbait Detection Review:

Rakhmawati and Zuhroh (2019) provide a review of clickbait detection methods by comparing traditional feature-based approaches with modern deep learning techniques. Their insights into effective feature extraction methods (e.g. word embeddings and sentiment features) could be incorporated into the detection module of my pipeline.

2.5 Fake News Classification:

Ramirez (2023) propose headline-based fake news detection using classic machine learning classifiers with features like tf-idf and sentiment scores. Their findings can be helpful for the classification module of my pipeline.

2.6 Generative Pretraining Frameworks:

Lewis et al. (2020) introduce BART, whose architecture supports my approach in two ways: its bidirectional encoder understands the full article context and its autoregressive decoder generates coherent text. Fine-tuning BART on headline-article pairs with corrective summaries to learn to identify misleading headlines and produce clarifying summaries that fill in omitted details seems promising.

2.7 Unified Text-to-Text Framework:

Raffel et al. (2020) propose T5, a framework that brings NLP problems into text-to-text format. I modify this to converting a headline-article pair into a corrective summary.

2.8 Abstractive Summarization with Pointer-Generator Networks:

See et al. (2017) improve sequence-to-sequence summarization with a pointer-generator network that can both copy from the source text and generate new tokens. I plan to use a pointer mechanism in my generative module to ensure that the corrective summaries actually reflect the key points from the article.

2.9 Pretraining for Summarization:

Zhang et al. (2020a) introduce PEGASUS, which pretrains models using a gap-sentence generation objective. I take up on this idea by adapting the pretraining objective to generate corrective summaries that fill in the gaps created by misleading headlines.

2.10 Distant and Weak Supervision:

Mintz et al. (2009) and Ratner et al. (2017) provide methods for distant supervision to automatically generate training labels from large unlabeled data corpora. I plan to potentially explore these techniques to supplement manual annotations, particularly for detecting subtle cues in misleading headlines.

2.11 Additional Inspirations:

I also considered works by Shu et al. (2017) and Pennycook et al. (2020) that examine fake news propagation and detection on social media. Even though their work is not directly used in my project, it shows the importance of explainability and user guidance in fighting misinformation. In addition, I reviewed evaluation metrics such as Zhang et al. (2020b), which leverage contextual embeddings for a more robust evaluation of generated text.

3 Datasets

After examining more than 10 datasets, I have selected a combination of datasets that provide diverse news content and annotations for both detection and generation tasks. For links to the datasets, see section References.

3.1 LIAR Dataset

The LIAR dataset consists of approximately 12,000 political statements labeled with truth ratings (half-true, false, mostly-true, true, barely-true, pants-fire). For the binary fake news classification task, I map the truth ratings as follows:

- Misleading: false, barely-true, pants-fire
- Reliable: half-true, mostly-true, true

This mapping results in the training of a binary classifier that distinguishes between misleading and reliable headlines.

3.2 Clickbait Challenge 2017 Dataset

This dataset contains annotated clickbait and non-clickbait headlines paired with the full articles. It is used to train and evaluate the headline detection system.

3.3 Cornell Newsroom Dataset

This large-scale dataset contains over one million news articles together with their headlines and summaries. It is used for training the corrective context generation module, ensuring that the generated

Dataset	Vars	Count	Words	Vocab	Dups
LIAR	14	10,239	18	21,676	17
Clickbait	2	32,000	9	35,789	0
Cornell	12	995,041	558	1,386,391	22
News Agg	8	422,419	9	176,490	15,964

Table 1: Descriptive statistics for the datasets used in the project. “Vars” is the number of variables (columns), “Count” is the document count, “Words” is the average word count per document, “Vocab” is the vocabulary size, and “Dups” is the duplicate count. For Cornell, the last three statistics are only for the first 100,000 documents, as the file was too large to explore it entirely.

summaries accurately capture the missing context from the full article content by comparing them to the given summaries.

3.4 News Aggregator Dataset

This dataset includes news headlines, full articles, and category labels from multiple sources. It will be used as a supplementary resource for evaluation and testing, ensuring that the overall pipeline generalizes well across diverse news sources in both classification and summarization tasks.

3.5 Excluded Datasets

- **Colossal Clean Crawled Corpus (C4):** Lacks targeted annotations
- **AllTheNews:** Too broad in scope
- **Freebase-derived Data:** Meant for relation extraction
- **MIND:** Focuses on click-through data
- **Reuters-21578:** Oriented toward topic classification
- **NELA Dataset:** Has heterogeneous annotations

4 Preliminary Experiments

At this stage, my primary objectives were to test the feasibility of my approach and refine the methods before full integration. I conducted the following experiments:

4.1 Detection Baseline:

I used the LIAR and Clickbait Challenge datasets. I extracted simple features such as tf-idf values and sentiment scores from the headlines and then trained a basic binary classifier based on Logistic Regression to determine if it could distinguish misleading headlines from reliable ones. The experiments on the LIAR dataset showed solid performance with moderate accuracy, while results on the clickbait dataset were surprisingly high—even though the data was balanced with only five duplicates between train and test—and the classifier achieved roughly 95% accuracy. The evaluation was focused on basic metrics (accuracy, precision, recall) along with a quick error analysis, and I plan to validate the model on external datasets.

4.2 Generation Baseline:

For the generation baseline, I experimented with several pre-trained summarization models, including facebook/bart-large-cnn, t5-small, and google/pegasus-xsum, using a small subset of the Newsroom dataset. I evaluated different summarization parameters (e.g., varying max and min lengths) and computed ROUGE scores to measure the similarity with the human-written summaries. Facebook/bart-large-cnn, particularly with a higher max_length (200) and min_length (50), achieved the best performance with ROUGE-1 F1 scores around 0.30 and ROUGE-2 F1 scores near 0.17. These ROUGE values suggest moderate overlap, but the evaluation was based on only 5 examples, and further validation on a larger set will be necessary. Overall, these findings seem to provide a promising starting point for further optimization and fine-tuning of the summarization module.

5 Evaluation Strategy

I will use a diverse evaluation strategy to assess the performance of my system:

5.1 Detection Evaluation

- I will use standard classification metrics (accuracy, precision, recall, F1 score) on held-out test sets from LIAR and the Clickbait Challenge dataset.
- I will perform an error analysis to identify common failure cases and refine the feature set.

5.2 Generation Evaluation

- I will evaluate the generated corrective summaries using ROUGE (RG-1, RG-2, RG-L) and BERTScore.

- I plan to conduct a small-scale human evaluation (using surveys or expert reviews) to assess clarity, correctness, and usefulness.

5.3 Optional: Distant Supervision Evaluation

- If distant supervision is implemented, I will measure the correlation between the detection model's attention outputs and a manually annotated set of key article segments.

- I will conduct ablation experiments comparing performance with and without distant supervision signals.

5.4 Reproducibility

- I will ensure that all code and data preprocessing scripts are well documented and available in a public repository.

- I will provide end-to-end scripts that allow others to replicate my experimental results.

6 Task Breakdown and Timeline

Given the deadline in seven weeks for the project (starting on March 24), I have planned the following timeline:

Week 1 (Mar 24 – Mar 30)

- Finalize the literature review and select key datasets
- Download and preprocess the LIAR, Clickbait Challenge, Reuters-21578, News Aggregator, and Cornell Newsroom datasets
- Set up the project repository and code structure
- Conduct preliminary experiments
- Create a timeline
- Write mid-term report

Deadline: Mar 30

Week 2 (Mar 31 – Apr 6)

- Implement the baseline detection module
- Train, evaluate and finetune a binary classifier using the LIAR and Clickbait Challenge datasets

Deadline: Apr 6

Week 3 (Apr 7 – Apr 13)

- Implement the baseline generation module by fine-tuning a pointer-generator/T5/BART model on the Cornell Newsroom dataset
- Generate initial corrective summaries and evaluate them using ROUGE/BERTScore

Deadline: Apr 13

Week 4 (Apr 14 – Apr 20)

- Conduct experiments on distant supervision by analyzing attention weights and feature importance
- Compare these signals against a small set of manually annotated key segments

Deadline: Apr 19

Week 5 (Apr 21 – Apr 27)

- Integrate the detection and generation modules into a single pipeline
- Perform end-to-end tests on a small batch of news articles and refine the approach

Deadline: Apr 26

Week 6 (Apr 28 – May 4)

- Refine experiments and evaluation strategies based on initial results
- Run the model on the entire datasets
- Write the final report and prepare the code for reproducibility

Deadline: May 4

Week 7 (May 5 – May 12)

- Buffer: Finish everything that is still unfinished
- Conduct final human evaluation studies

Deadline: May 12

7 Conclusion

In this midterm report, I have demonstrated my progress on a system to automatically generate corrective context for misleading news headlines. I have completed a thorough literature review, selected, accessed, explored and preprocessed several key datasets, and conducted preliminary experiments to validate both the detection and generation modules. My evaluation strategy combines automatic metrics with human judgment to assess performance in both components. The task breakdown and timeline provide clear milestones for the remaining weeks. The next steps involve executing the preliminary experiments, refining my methods based on the results, and integrating all components into a reproducible end-to-end system.

Limitations

At this stage, my approach is limited by computational resources and the fact that I am working alone. My initial experiments use only a subset of the data. I expect to adjust the methods as more experimental results are gathered.

References

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL 2020*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL 2009*, pages 1003–1011.

Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Management Science*, 66(11).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nur Aini Rakhmawati and Nurrida Aini Zuhroh. 2019. Clickbait detection: A literature review of the methods used. In *Jurnal Ilmiah Teknologi Sistem Informasi*, volume 6.

Gared Ramirez. 2023. Fake news detection with headlines.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Re. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of VLDB Endowment*, volume 11, pages 325–338.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL 2017*, pages 1073–1083.

Robert Sepulveda Torres, Alba Bonet-Jover, and Estela Saquete. 2023. Detecting misleading headlines through the automatic recognition of contradiction in spanish. In *IEEE Access*, volume 99. Extended abstract available on ResearchGate.

Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Michael Bendersky, and Marc Najork. 2023. "why is this misleading?": Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, pages 1662–1672. ACM.

Kai Shu, Anna Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. In *ACM SIGKDD Explorations Newsletter*, volume 19, pages 22–36.

Kai Shu, Anna Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2019. defend: Explainable fake news detection. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pages 396–405. Extended version focusing on explainability aspects.

Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2022. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of AAAI*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of ICML 2020*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR 2020*.

Datasets

The following datasets are used in this work:

- **LIAR Dataset:** Approximately 12,000 political statements with truth ratings and metadata. Available on Kaggle at https://www.kaggle.com/datasets/doanquanvietnamca/liar-dataset.
- **Clickbait Challenge 2017 Dataset:** Annotated clickbait and non-clickbait headlines along with full articles. Available on Kaggle at https://www.kaggle.com/datasets/amananandrai/clickbait-dataset.
- **News Aggregator Dataset:** News headlines, full articles, and category labels from multiple sources. Available on Kaggle at https://www.kaggle.com/datasets/uciml/news-aggregator-dataset.
- **Cornell Newsroom Dataset:** Over one million news articles paired with headlines. Available at https://lil.nlp.cornell.edu/newsroom/index.html.

A Appendix

Link to the Github Repository: https://github.com/row56/headlinecontext

Additional details, hyperparameter settings, and extended experimental results will be included in the final report.