De-clickbaiting the News

By: Arjun Karanam, Ronak Malde, Michael Elabd

CS 224n

Abstract

News headlines have long been criticized for being "clickbait", where the headline misleads the reader to draw their attention. While there have been various solutions to classify clickbait headlines and generating a new headline from scratch, little has been done to both preserve the author's intent while ensuring clarity of information. In this project, we used a pretrained encoder-decoder model, T5, and proposed two novel methods to achieve these goals. First, a custom loss function based on a pretrained BART summarizer, and second a loss penalty based on a BERT-based finteuned clickbait classifier. We found that

Problem and Motivation

News headlines have long been criticized for being "clickbait", where the headline misleads the reader to draw their attention. While there have been various solutions to classify clickbait headlines and generating a new headline from scratch, little has been done to both preserve the author's intent while ensuring clarity of information. The goal for this project is to modify clickbait news headlines such that the new headline is more representative of the article, while also preserving elements of the original headline.

Data

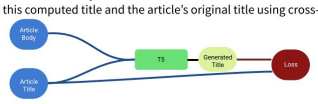
All models were trained using the Webis-Clickbait-17 dataset. The dataset contains a total of 40976 labelled articles along with 80013 unlabelled articles. These articles are published between November 2015 and June 2017 and are only from United States news outlets. Each article is bucketed into 4 categories of the headline's "clickbait score", judged by human reviewers. Webis defined clickbait as a headline that is designed to entice its readers into clicking an accompanying link i.e. something unnamed is referred to, some emotional reaction is promised, some lack of knowledge is ascribed, some authority is claimed, etc. We used this dataset to finetune the BERT classifier and to train the T5 model.

Methods

We wanted to generate headlines that did these three things: Summarized the original article, retain the author's original intent, and was overall not clickbait. For each goal, we came up with a unique loss metric.

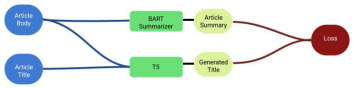
Retain Author Intent

Most headline generators today look at the article body, and generate a headline. We see value in retaining the author's original intent, as it may include information or a style that can't be inferred from just the body. To accomplish this, we take both the article body and article title and use a fine-tuned T5 model to generate a title. Then we compute a loss between this computed title and the article's original title using cross-entropy loss.



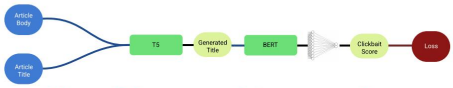
Summarized Article

First and foremost, a good headline should give an indication what an article is about, while not misrepresenting the article's contents. To achieve this, we design a loss that compares a generated title to the article itself. To accomplish this, we again use both the original article and its title to generate a title using a fine-tuned T5 model. Additionally, we use a BART summarizer to generate an article summer given the article body. Finally, we use cross-entropy loss to compoare the generated title to the generated article summary.

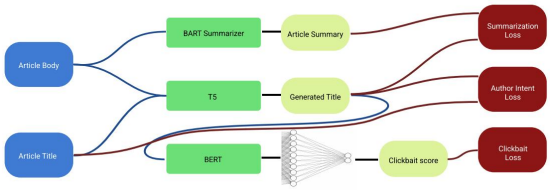


Avoid Clickbait

As we discuss in our data section, it's hard to precisely define what is and isn't clickabit. From our dataset, we train a classifier that is able to distinguish clickbait from non clickbait titles, and use that to design a penalty. Like before, we use a fine-tuned T5 model to generate article titles given the original article title and body. Additionally, we train a BERT classifier on the aforementioned Webis dataset to classify clickbait vs. non clickbait headlines. Finally, we take our generated headline, classify as clickabit or non-clickbait using our BERT classifier, and compute a penalty if the headline reads like clickbait.



Algorithms and Approaches



Results

Train Dataset:

Method	BLEU	Cosine Sim (Spacy)	Cosine Sim (SBERT)	Rouge 1	Rouge 2
T5: Author Intent Loss	0.026	0.57	0.27	0.033	0.0060
T5: Author Intent Loss + Summarization Loss	0.027	0.62	0.28	0.038	0.0067
T5: Author Intent Loss + Classification Penalty	**0.029**	**0.63**	**0.31**	**0.044**	**0.0084**
T5: Author Intent Loss + Summarization Loss + Classification Penalty	0.025	0.62	0.28	0.037	0.0058

Test Dataset:

Method	BLEU	Cosine Sim (Spacy)	Cosine Sim (SBERT)	Rouge 1	Rouge 2
T5: Author Intent Loss	0.023	0.575	0.282	0.035	0.0077
T5: Author Intent Loss + Summarization Loss	0.025	**0.647**	0.297	0.040	0.0068
T5: Author Intent Loss + Classification Penalty	**0.034**	0.642	**0.306**	**0.041**	**0.008**
T5: Author Intent Loss + Summarization Loss + Classification Penalty	0.024	0.641	0.243	.0267	0.004

Analysis

Both the summarization loss and the classification penalty improved the T5 model's scores generally on the train and test datasets. Most notably, there was great improvement in the cosine similarities for Spacy and SBERT embeddings with both methods with improvements around .05, meaning that that the title retained semantic meaning better. Additionally, the two methods performed better on Rouge scores, better summarizing the article itself. When combining the two methods, however, the model did far worse on the test set than the models invidiually. It recieved a particularly low score on the Rouge 2 metric, almost half of the base model.

Example output for summarization loss and classification penalty:
Original title - "Boat Race: unexploded bomb found near starting line"
Generated title - "police were called after a member of the public spotted second world war bomb near Putney"

Future Work

While these initial results are promising for the area of unclickbait headlines, there is still a lot of room to improve our existing architectures. We can further tune our hyperparameters like learning rate, AdamW optimizer parameters, and weight decay, using parameter grid search. Additionally, we would like to make our metrics more robust to assess how well the headline achieves our goals. ROUGE scores seem inadequate, and it might be useful to create a large-scale human scoring system, similar to how the dataset was constructed. This could also help identify specific edge cases with outputs.
One future area for these methods could be style transfer of generated headlines, as an abstraction of our current approach. This could be used to not only unclickbait a headline, but to make a headline more funny, or more academic.