Fine-Tuning FLAN-T5 For Spoiling Clickbait Content
NLP Course Project & Project Work

Michael Ghaly, Master's Degree in Artificial Intelligence, University of Bologna
michael.magdynasr@studio.unibo.it

Abstract

This report presents the project's objective, approach, and observations in the context of Task 2 of the Clickbait Challenge at SemEval 2023. The aim is to generate informative spoilers for clickbait posts by fine-tuning the FLAN-T5 transformer model solely on the posts' clickbait title, content and human-generated spoilers provided in the dataset. The approach harnesses the transformer's sequence-to-sequence architecture to generate coherent and informative spoilers. The findings underscore the transformer's efficacy in producing contextually relevant responses, showcasing its potential in enhancing clickbait spoiling techniques. This report provides insights into the effectiveness of transformer-based models for generating compelling content that piques readers' curiosity while adhering to the challenge's objectives.

1 Introduction

In the era of digital content consumption, clickbait posts have become ubiquitous, enticing users with curiosity-driven headlines that often fall short of delivering substantive content. This practice not only misleads users but also affects the credibility of online platforms. The Clickbait Challenge 2023 addresses this issue by generating informative spoilers for clickbait posts, satisfying the curiosity induced by a clickbait post. Such a system aligns with ethical content sharing but also enhances user experience.

Standard clickbait mitigation approaches include sentiment analysis, keyword matching, and machine learning models. However, these methods often focus on flagging clickbait posts rather than providing informative content beforehand. More advanced systems rely on sequence-to-sequence models like transformers to generate content summaries. This project, inspired by recent advancements in transformer-based natural language processing, addresses the challenge by fine-tuning the FLAN-T5 base transformer model exclusively on the clickbait post's title and content, leveraging the transformer's ability to capture contextual nuances and generate coherent and informative spoilers that align with the challenge's objectives.

The results underscore the effectiveness of the transformer-based approach in generating informative spoilers. The method showcases the potential of transformer models to provide contextually relevant content that enriches user experience. By contributing insights into the capabilities of such models, this work advances the field of clickbait spoiling and offers a promising avenue for creating engaging, ethical content summaries.

2 Background

Clickbait content refers to online articles, headlines, or posts employing intriguing language to attract attention and entice users into clicking on them. Often, they overpromise and underdeliver, misleading readers about the actual content. This practice not only compromises user trust, but also negatively impacts the quality of information dissemination. Clickbait's prevalence undermines the credibility of online platforms and disrupts users' experiences by leaving them unsatisfied after engaging with the content.

Sequence-to-Sequence architectures are a class of deep learning models designed for natural language processing tasks. These architectures consist of two main components: an encoder and a decoder. The encoder encodes the input sequence into a fixed-length representation, often called a "context vector." The decoder then generates an output sequence based on this context vector. This architecture is particularly well-suited for tasks like machine translation, text summarization, and conditional content generation as they excel in capturing the underlying structure and relationships within sequences, allowing them to handle complex language tasks effectively. A variant of seq2seq models, known as transformer-based architectures, has gained significant attention in recent years. Transformers leverage self-attention mechanisms to process input sequences in parallel and capture long-range dependencies. Pre-trained transformer models, such as BERT, GPT, and T5, have demonstrated remarkable capabilities across various natural language processing tasks due to their capacity to understand context and generate coherent text.

FLAN-T5 stands for "Fine-tuned LAnguage Net" and "Text-To-Text Transfer Transformer," respectively. The origins of FLAN-T5 can be traced back to Google's work in (Raffel et al., 2020) when they introduced the original T5 architecture. T5 demonstrated its prowess in various tasks, particularly in translation and summarization. Building upon this foundation, Google further refined the model in 2022 through (Chung et al., 2022). This evolution resulted in FLAN-T5, a language model that achieved strong few-shot performance even compared to much larger models. At its core, FLAN-T5 is an encoder-decoder model. During its training, the model has been exposed to a diverse range of tasks phrased as instructions rather than being specialized in a single task, which enables models to respond better to instructions and reduces the need for few-shot exemplars. The paper found that this technique, called instruction finetuning, coupled with scaling the number of tasks and scaling the model size dramatically improves performance on a variety of model classes (PaLM, T5, U-PaLM), prompting setups (zero-shot, few-shot, CoT), and evaluation benchmarks.

Applying FLAN-T5 to Clickbait Spoiling

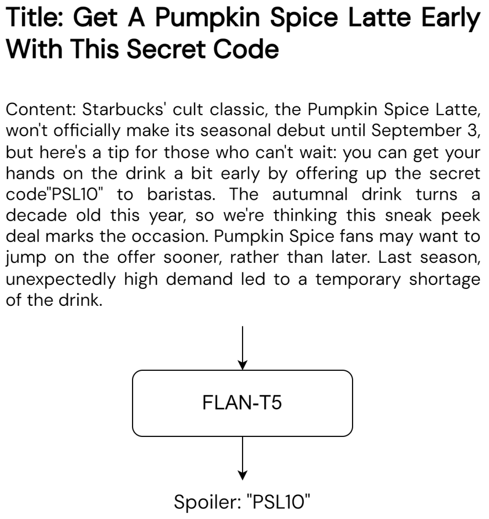
This project finetunes the base version of the FLAN-T5 model to tackle the clickbait content problem. The challenge's dataset, encompassing clickbait posts and human-generated spoilers, serves as the foundation for training and evaluating the model to capture the essence of the linked content and generate a spoiler that satisfies the curiosity induced by the clickbait headline.

3 System description

The implemented system for addressing Task 2 of the Clickbait Challenge involves fine-tuning FLAN-T5 base, a encoder-decoder transformer architecture, to generate informative spoilers for clickbait posts. The core components and processes of the system include:

1. **Data Preprocessing** The clickbait challenge dataset, consisting of 3,000 training samples, 200 validation samples, and 800 test samples, undergoes preprocessing. Each clickbait post's title and content are extracted to form input sequences. A training sample is shown as follow:

Figure 1: Training Instance



Title: Get A Pumpkin Spice Latte Early With This Secret Code

Content: Starbucks' cult classic, the Pumpkin Spice Latte, won't officially make its seasonal debut until September 3, but here's a tip for those who can't wait: you can get your hands on the drink a bit early by offering up the secret code"PSL10" to baristas. The autumnal drink turns a decade old this year, so we're thinking this sneak peek deal marks the occasion. Pumpkin Spice fans may want to jump on the offer sooner, rather than later. Last season, unexpectedly high demand led to a temporary shortage of the drink.

Spoiler: "PSL10"

2. **Fine-Tuning** The FLAN-T5 base transformer model is fine-tuned on the tokenized training dataset for five epochs. During fine-tuning, the model learns to map the input sequences (clickbait post titles and content) to corresponding informative spoilers.

3. **Evaluation** The fine-tuned model's performance is evaluated on the validation dataset, utilizing the BERTScore as the primary evaluation metric. In contrast to n-gram-based metrics like BLEU and ROUGE, commonly used for text generation tasks, the BERTScore metric leverages pretrained BERT contextual embeddings (Devlin et al., 2019). BERTScore (Zhang et al., 2020) assesses sentence similarity by calculating the sum of cosine similarities between embeddings of corresponding tokens. This approach addresses inherent limitations in n-gram-based metrics. For instance, n-gram metrics fail to robustly match paraphrases, leading to underestimation of performance when semantically-correct phrases deviate from the reference's surface form. BERTScore's contextualized token embeddings effectively tackle these pitfalls.

They capture contextual nuances and have proven effective in paraphrase detection, ensuring that semantically equivalent phrases are appropriately recognized. Additionally, BERTScore excels at capturing distant dependencies and critical ordering changes, often inadequately addressed by n-gram models. This makes it well-suited for evaluating the quality of informative spoilers, where the essence and coherence of generated content are crucial.

4. **Inference** After fine-tuning, the model is uploaded to the Hugging Face hub. It can be downloaded at any point and used for inference. Given a clickbait post, the model generates an informative spoiler that concisely captures the essence of the linked content.

4 Data

The Webis Clickbait Spoiling Corpus contains 5,000 spoiled clickbait posts crawled from Facebook, Reddit, and Twitter. This dataset contains the clickbait posts and manually cleaned versions of the linked documents, and extracted spoilers for each clickbait post. Additionally, the spoilers are categorized into three types: short phrase spoilers, longer passage spoilers, and multiple nonconsecutive pieces of text.

The dataset comes with predefined train/validation/test splits:

- 3,200 posts for training
- 800 posts for validation
- 1,000 posts for testing (omitted from the dataset)

Thus, for development, the training samples are split into 3,000 for training and 200 for validation. The 800 validation samples are used as a test set.

5 Results and Discussion

The performance of the implemented system was qualitatively evaluated by inspecting the generated predictions and comparing them to the ground truth spoilers provided in the test set. The model's results demonstrate promise in generating informative spoilers for clickbait posts.

Table 1: Comparison of Predictions and Ground Truths

Title: Here's how much you should be tipping your hairdresser **Prediction:** 20% **Ground Truth:** 20%
Title: Man swallowed a microSD card and you won't believe what happened next! **Prediction:** pooped it into a strainer **Ground Truth:** a man who swallowed a 64GB microSD card and then pooped it into a strainer
Title: The Anytime Snack You Won't Feel Guilty About Eating (VIDEO) **Prediction:** Smoky Paprika-Baked Garbanzo Beans **Ground Truth:** Smoky Paprika-Baked Garbanzo Beans
Title: Do birds fart? **Prediction:** birds are capable of passing gas, but they don't because they lack the same combination of diet and gas-forming bacteria in their gut that gives mammals their odorous gas. **Ground Truth:** They don't fart

Upon manual inspection of a diverse subset of predictions, it was observed that the generated spoilers consistently captured the essence of the linked content. In certain cases, the model responses were too concise and too focused on individual names. Shown below are some instances that illustrate this limitation: This issue could potentially be ameliorated by adjusting the preprocessing of the human-annotated spoilers and tweaking the generation hyperparameters.

Table 2: Comparison of Predictions and Ground Truths

Title: Why Arizona Sheriff Joe Arpaio's fate could hang on a single word **Prediction:** "intentionally" **Ground Truth:** "intentionally" - could transform a court case against Phoenix-area Sheriff Joe Arpaio from civil charges to a criminal prosecution
Title: Local Cops Make SHOCKING Move Against Boy Selling Teddy Bear To Buy Food **Prediction:** Officer Steve Dunham **Ground Truth:** Dunham picked the boy up and took him to a Subway to get something to eat. He then took him to the Franklin Police Department.

6 Conclusion

This project addressed Task 2 of the Clickbait Challenge 2023 by fine-tuning the sequence-to-sequence FLAN-T5 transformer model to generate informative spoilers for clickbait posts. The approach showcased promising results, aligning with the challenge's objectives and demonstrating the potential of transformer models in content generation. Through qualitative evaluation, it was observed that the generated spoilers consistently encapsulated the key themes of linked articles or web pages. The model's outputs retained contextual nuances, displayed coherence, and closely resembled human-generated spoilers. This success was consistent across diverse clickbait post lengths and complexities, validating the versatility and robustness of the model. While the results were largely in line with expectations, there were some instances where the model struggled to capture specific nuances, leading to slight deviations from the ground truth spoilers. Despite these nuances, the overall performance was noteworthy, and the model's ability to generate informative spoilers for a range of clickbait posts was promising. The main limitation of our solution lies in its qualitative evaluation approach, which relies on human inspection and subjective judgments. Future work could benefit from more quantitative evaluations and a larger dataset to ensure a comprehensive training and assessment of the model's performance. Overall, our work highlights the potential of transformer-based models in enhancing clickbait spoiling techniques and paves the way for future advancements in this domain.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.