Clickbait Content Spoiler Classification and Generation
NLP Course Project

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

Abstract

This report presents the objectives, methodologies, and observations of my submission for the Clickbait Challenge at SemEval 2023. The primary goal of this work is to generate informative spoilers for clickbait posts by leveraging transformer models. The submission consists of two tasks. For Task 1, the focus is on classifying the spoiler types associated with clickbait posts by fine-tuning RoBERTa on the multi-label classification task. While Task 2 entails the generation of coherent and meaningful spoilers 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 today's 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. Addressing this issue, the Clickbait Challenge at SemEval 2023 tackles the task of classifying clickbait posts and generating informative spoilers aimed at satiating the curiosity sparked by them. Such an approach not only aligns with ethical content dissemination but also elevates the overall 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 solutions leverage sequence-to-sequence models such as transformers to generate concise summaries. This project, inspired by recent advancements in transformer-based natural language processing, addresses the challenge by fine-tuning transformer models (RoBERTa for task 1 and FLAN-T5 for task 2) exclusively on the clickbait post's title and content, leveraging the transformer's ability to capture contextual nuances to classify them 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

2.1 Clickbait Spoiling

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.

2.1.1 Task 1: Spoiler Type Classification

Task 1 of the challenge focuses on Spoiler Type Classification. The objective is to predict the appro-

priate spoiler type for a given clickbait post and its linked document. The spoiler types are categorized into three classes: "phrase," "passage," and "multi." Each of these types corresponds to a specific level of detail and length in the provided spoiler. Accurate classification of the spoiler type is crucial for generating informative summaries that align with user expectations.

Spoiler type classification presents several challenges rooted in the nuances of clickbait content. Differentiating between spoiler types requires understanding the granularity of the information revealed in the spoiler and its alignment with the user's curiosity. Additionally, the presence of linguistic ambiguity and context-dependence within clickbait posts and linked documents adds complexity to the classification task. Balancing the need for providing enough detail to be informative while avoiding over-disclosure poses an intricate challenge for NLP models.

2.1.2 Task 2: Spoiler Generation

Clickbait's allure lies in its ability to invoke curiosity and compel users to click on links, boosting engagement metrics for online platforms. However, this strategy frequently leads to a dissonance between user expectations and actual content. Users may find the content to be underwhelming or misrepresentative, which can ultimately erode trust and engagement.

Clickbait spoiling aims to bridge the gap between user expectations and actual content by offering informative concise summaries. By generating spoilers that accurately encapsulate the main content, users can make informed decisions about whether to engage further. Spoiling addresses the need for greater transparency and user satisfaction while still retaining a degree of intrigue that encourages engagement.

2.2 RoBERTa

A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019) stands as a significant milestone in the realm of pre-trained language models. Building upon the foundation of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), RoBERTa was designed to enhance language understanding and performance across diverse natural language processing tasks. Developed by Facebook AI in 2019, RoBERTa addresses limitations in BERT's pre-training methodology by leveraging a larger corpus, dynamic masking, and unidirectional training. These refinements have led to its remarkable capacity to capture nuanced linguistic relationships and context, making it a potent tool for tasks ranging from sentiment analysis to machine translation. In the context of the Clickbait Challenge, RoBERTa's contextualized embeddings and fine-tuning capabilities offer a compelling approach to tackle the Spoiler Type Classification task, providing accurate classifications of spoiler types for clickbait posts and linked documents.

2.3 Sequence-to-Sequence Architecture

Seq2Seq models 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.

2.4 FLAN-T5

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 spe-

cialized 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.

3 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 non-consecutive 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.

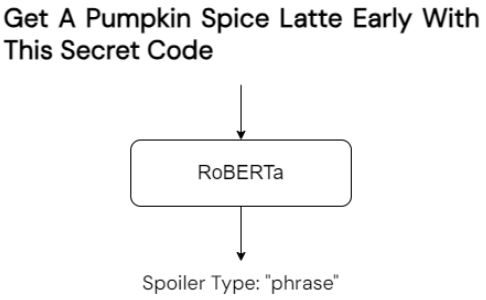
4 Task 1 System Description

The implemented system for addressing Task 2 of the Clickbait Challenge involves fine-tuning distil-RoBERTa, an encoder-only transformer, as a classifier for the spoiler type warranted by a clickbait post's title. Thus, the model is trained solely on the headline of the post and not the its content.

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 is extracted to form input sequences. A training sample is shown as follow:

   An initial experiment was done using the content of the post as well but the difference in the model's prediction accuracy was insignificant.

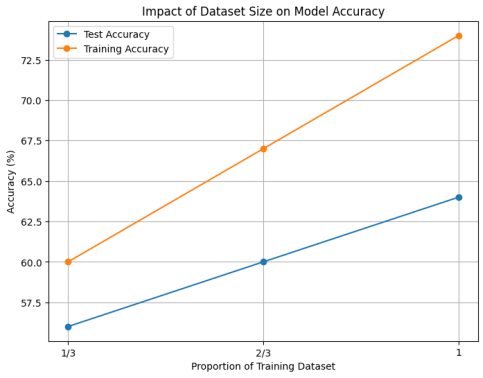
Figure 1: Training Instance



2. **Fine-Tuning and Evaluation** The distil-RoBERTa 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) to the corresponding spoiler type. For the evaluation, a simple prediction accuracy is deemed a sufficient indicator to the performance of the model.

3. **Results and Discussion** The classifier model achieved a final accuracy of 62% on the test set and 74% on the training set. While these accuracy values might not be exceptionally high, they provide valuable information about the model's capabilities and potential areas for improvement. It's plausible to attribute this to the constraints imposed by the small dataset. To further investigate this suspicion, a systematic exploration of the relationship between dataset size and model performance was undertaken. The model was subjected to a series of fine-tuning experiments using varying proportions of the training data: one-third, two-thirds, and the complete dataset.

Figure 2: Dataset Impact on Model Performance



The outcomes of these experiments consis-

tently reinforce the notion that the limited dataset size is indeed a contributor to the observed low accuracy. Additional factors can be attributed to the ambiguity and subjectivity of clickbait titles. In fact, the confusion matrix shows that the model stuggles much more with the "mutli" label compared to the other two. For example, the clickbait headline input: "Why Arizona Sheriff Joe Arpaio's fate could hang on a single word" was predicited to have the label "passage" as the spoiler should mention the word and explain the context, but it could also be a list of spoiler points containing the word and its explanation.

5 Task 2 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 3: Training Instance

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.

6 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:

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.

This issue could potentially be ameliorated by adjusting the pre-processing of the human-annotated spoilers and tweaking the generation hyperparameters.

7 Conclusion

This project addressed Task 1 and 2 of the Clickbait Challenge 2023 by fine-tuning the encoder distil-RoBERTa and the sequence-to-sequence FLAN-T5 transformer models to classify and 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 classification and 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 the proposed 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.