Clickbait Spoiling with Language Models

Berke Can Rizai
brizai18@ku.edu.tr

Eren Barış Bostancı
ebostanci18@ku.edu.tr

Abstract

This paper proposes an approach for the SemEval 2023 Task 5 challenge, which involves predicting spoilers in news articles. Spoiler writing is a topic that has been explored in various approaches from basic models to more advanced transformer models in recent years.

Previous research in this field has explored various approaches, including rule-based methods, machine-learning models, and deep-learning models. Some studies have focused on using encoders and question-answering models, while others have used reinforcement learning, headline generation with constraints of body text as well as GANs. However, these approaches may be outperformed by recent developments in LLMs as they have been proven successful in many downstream tasks including document question answering and generative search.

Our approach builds on recent advances in Large Language Models (LLMs), which have shown promising results in various natural language processing tasks. Specifically, we plan to use the StableLM, a state-of-the-art LLM, which we fine-tune using the provided dataset and additional data created using ChatGPT. We also explore other LLMs and implement baseline models using TF-IDF and Bert for question-answering.

Our primary focus is on one-shot or few-shot learning methods to minimize the need for large amounts of labeled data. This is particularly important for the SemEval Task 5 challenge, as the dataset is relatively small and may not capture the full range of spoilers in news articles. By leveraging the power of LLMs and few-shot learning methods, we aim to improve the accuracy and efficiency of spoiler prediction models.

We evaluate our approach using the SemEval Task 5 guidelines and metrics, including Meteor 1.5. Our approach has the potential to outperform previous methods and provide insights into the effectiveness of LLMs for spoiler detection. Our work has implications for various applications, such as content moderation, user privacy, and news recommendation systems.

1 Introduction

Clickbait has become a prevalent issue on the internet, often misleading users with sensational headlines to drive traffic to websites. In this project, we aim to tackle the problem of clickbait by generating spoilers for clickbait posts using SOTA Models (LLMs) and evaluate them with some other baseline models and recent approaches. We hypothesize that fine-tuning LLMs on this task can lead to high-quality spoilers that effectively neutralize the clickbait.

Our research questions include:

- Can LLMs be successfully fine-tuned for clickbait spoiling, and how well do they perform compared to other baseline methods?
- Can one-shot or few-shot learning methods be applied to this task to reduce the need for large amounts of labeled data?
- Can we utilize some recent developments in fine-tuning LLMs, such as LoRA for decreased computational cost and better performance?

2 Related Work

Variety of methods have been employed for clickbait spoiling, some methods such as style transfer has been partially successfull "we propose to generate stylized headlines from original documents with style transfer"(Shu et al., 2018) researchers propose a novel method called Stylized Headline Generation that an iteration of style transfer. Researcher implemented an autoencoder where they have used RNN as an encoder,

Another research by (Xu et al., 2019) proposes a reinforcement learning approach. Authors propose a novel approach for generating sensational headlines without relying on labeled data. The authors emphasize the unique characteristics of headlines, which are meant to capture readers' attention and generate interest. They identify the challenge of lacking a sensationalism scorer for headline generation and introduce a distant supervision strategy to collect a sensationalism dataset for training the scorer. This strategy involves using headlines with high comment counts as positive samples, and generating headlines from a summarization model as negative samples. Authors generate negative samples by utilizing a summarization model. They use the summarization model to generate summaries of the articles, and then extract the generated summaries as negative samples. Since the generated summaries are not expected to contain spoilers, they serve as negative examples for training the model to differentiate between spoilers and non-spoilers.

The retrieved spoilers from the positive samples and the generated summaries from the summarization model are combined to create a dataset for training the sensationalism scorer. The sensationalism scorer is a binary classifier that is trained to classify headlines as either sensational (containing spoilers) or non-sensational. The dataset created from the distant supervision strategy is used to train the sensationalism scorer, which is then used as a reward model in the reinforcement learning (RL) process for headline generation.

Another important contribution of the paper is addressing the issue of noisy rewards in reinforcement learning (RL). The authors introduce a new loss function called Auto-tuned Reinforcement Learning (ARL) to mitigate this challenge. They highlight that maximizing the sensationalism score from the sensationalism scorer alone may result in unnatural phrases, as the scorer may make mistakes and RL can generate sentences that artificially boost the score. To overcome this, the proposed ARL model automatically adjusts the balance between maximum likelihood estimation (MLE) and RL during training based on the sensational intensity of the training headline.

(He et al., 2020) Proposes DeBERTA which is a masked language model that particularly performs well on question-answering tasks of SQUAD and SuperGLUE. "For a token at position i in a sequence, we represent it using two vectors, which represent its content and relative position with the token at position j, respectively. The calculation of the cross attention score between tokens i and j can be decomposed into four components a sum of four attention scores using disentangled matrices on their contents and positions as content-to-content, content-to-position, position-to-content, and position-to-position 2." Authors also state that in addition to BERT's usage of positions, DeBERTa makes few adjustments.

For example, in a sentence like "a new store opened beside the new mall", where the words "store" and "mall" are masked for prediction, relying solely on local context (e.g., relative positions and surrounding words) may not be enough for the model to distinguish between "store" and "mall" since they have the same relative positions after the word "new". To address this limitation, DeBERTa incorporates absolute positions as complementary information to relative positions. This is done right after all the Transformer layers but before the softmax layer for masked token prediction, and it is called an Enhanced Mask Decoder (EMD). This way, DeBERTa captures relative positions in all the Transformer layers and uses absolute positions during the decoding process to improve prediction accuracy. In contrast, the BERT model incorporates absolute positions in the input layer. (He et al., 2020)

In the paper (Chakraborty et al., 2016), a 93% accuracy in detecting spoilers and an 89% accuracy in blocking them was achieved using common classification models like SVM, Decision Tree, and Random Forest. The authors used feature selection which included sentence structure, word patterns, clickbait language, and N-gram features. Sentence structure included headline length, stop word ratio, and average word length. Word patterns included the presence of cardinal numbers, unusual punctuation patterns, and the number of contracted word forms. Clickbait language captured the nuances of language used. N-gram features included Word N-grams, POS N-grams, and Syntactic N-grams. The paper also provided a benchmark for comparing each model using each feature and using all features. The best results were obtained with SVM using all features.

In their article (Karn et al., 2019), experiments with two state-of-the-art neural abstractive summarization techniques, attentive seq2seq, and pointer

seq2seq, for teaser generation and provides a benchmark. The attentive seq2seq model is designed to generate a target using words from a fixed vocabulary and the pointer seq2seq model uses a flexible vocabulary that is expanded with words from the source via the pointer network. The models are evaluated using Rouge-1, Rouge-2, and Rouge-L metrics. The results show that seq2seq point better than seq2seq due to the boost in recall gained by copying source words through the pointer network.

At the evaluation phase, human written spoilers are compared with retrieved answer, researchers referenced BLEU score (Papineni et al., 2002) for the implementation of such scoring.

We will be utilizing either of the two most popular methods for fine-tuning. First one is retraining all parameters on new data and second one which is more promising is LoRA. From the paper, "LoRA reduces the number of trainable parameters by learning pairs of rank-decomposition matrices while freezing the original weights. This vastly reduces the storage requirement for large language models adapted to specific tasks and enables efficient task-switching during deployment all without introducing inference latency. LoRA also outperforms several other adaptation methods including adapter, prefix-tuning, and fine-tuning." (Hu et al., 2021).

3 Model

We plan to use the StableLM, a state-of-the-art Large Language Model. We will fine-tune this model using the data provided in SemEval Task 5 and additional data, which we will create using ChatGPT. We will also consider other Large Language Models from this document. Our primary focus will be on one-shot or few-shot learning methods to minimize the need for large amounts of labeled data. We will also implement a baseline model using TF-IDF for question answering and test Bert for question answering as an alternative baseline.

4 Experimental Setup for Our Approach

- Dataset: SemEval Task 5 dataset and additional data created using ChatGPT
- Models: StableLM (fine-tuned), LLAMA, TF-IDF baseline, Bert for question answering
- Evaluation Metrics: Evaluation metrics defined by the SemEval Task 5 guidelines
- Approach: Fine-tune LLMs with various fine-tuning strategies and give the task with one shot or few shot examples to different LLMs.

4.1 Dataset

The Task 5's dataset contains the manually cleaned versions of this (Hagen et al., 2022) and extracted spoilers for each post. The spoilers are categorized into three classes, which are Short Phrase Spoilers, Longer Passage Spoilers, and Multiple Non-Consecutive Pieces of Text. The dataset is already divided. 3200 posts are available for training and 800 posts are available for validation. We plan to fine-tune the model with the training data and evaluate it with the validation data. We will also create additional labeled data by scraping more news and we will label them by using ChatGPT to augment the training set and potentially improve the performance of our model.

Format of the data in JSON and contains uuid, postText, targetParagraphs, targetTitle, targetUrl, humanSpoiler, spoiler, spoilerPosition, tags, and some metainformations such as postId, postPlatform, targetDescription... We will probably only use targetParagraphs, targetTitle, targetUrl, humanSpoiler, and spoiler columns.

The output formats are identical for part 1 and part 2. Both will contain uuid, spoiler-type ad spoiler values.

4.2 Baseline Models

We will implement two baseline models for comparison with our fine-tuned StableLM. The first baseline model will use TF-IDF for question answering based on sentence similarities to the headline. The second baseline will use Bert for question answering. These baselines will help us evaluate the effectiveness of our LLM-based approach. For the TF-IDF and BERT, one possibility is to approach the problem as question answering problem where the question is the non-spoiled click-bait title of the article and the target (answer) is a spoiler line from the article. Likewise, we will use the BERT as an encoder for the same task with the same approach.

4.3 Evaluation Metrics

We will use the evaluation metrics defined by the SemEval Task 5 guidelines to measure the performance of our models. Even though there is no description for the 2023 challenge, one of the

metrics is Meteor 1.5: Automatic Machine Translation Evaluation System. This (Denkowski and Lavie, 2011) is basically looking for how much does paraphrase of the spoiler (as annotated) and the retrieved prediction from the model correlate or how similar are they.

5 Schedule

Subtask	Assigned to	Time Estimate
Make research about possible LLMs, LoRA, and finetuning	Berke & Eren	1 week
Acquire and preprocess data	Berke & Eren	2 weeks
Fine-tune StableLM	Berke & Eren	4 weeks
Implement baseline models (TF-IDF)	Berke	1 week
Implement baseline models (Bert)	Eren	1 week
Analyze the output of the models	Berke & Eren	1 week
Work on the final report	Berke & Eren	1 week

Table 1: Project Timeline and Subtask Assignments

References

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. 2022. Clickbait Spoiling via Question Answering and Passage Retrieval. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 7025–7036. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Sanjeev Kumar Karn, Mark Buckley, Ulli Waltinger, and Hinrich Schütze. 2019. News article teaser tweets and how to generate them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3967–3977, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 467–476.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning.