SemEval-2023 Task 5: Clickbait Spoiling

Maik Fröbe
Friedrich-Schiller-Universität Jena

Tim Gollub
Bauhaus-Universität Weimar

Benno Stein
Bauhaus-Universität Weimar

Matthias Hagen
Friedrich-Schiller-Universität Jena

Martin Potthast
Leipzig University and ScaDS.AI

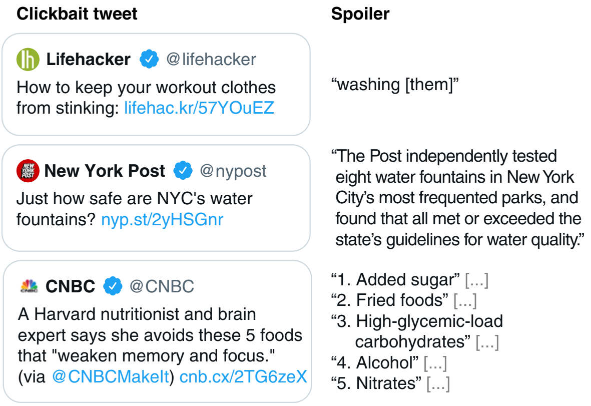
Abstract

In this overview paper, we report on the second PAN Clickbait Challenge hosted as Task 5 at SemEval 2023. The challenge's focus is to better support social media users by automatically generating short spoilers that close the curiosity gap induced by a clickbait post. We organized two subtasks: (1) spoiler type classification to assess what kind of spoiler a clickbait post warrants (e.g., a phrase), and (2) spoiler generation to generate an actual spoiler for a clickbait post.

1 Introduction

Clickbait posts link to web pages and advertise their content by arousing curiosity instead of providing informative headlines or summaries. The PAN Clickbait Challenge at SemEval 2023 aims to develop techniques for automatic clickbait spoiling: generating a short text that can close the curiosity gap as exemplified in Figure 1. Despite manual clickbait spoiling being popular on social networks,[1] the problem had not received much attention by the research community. We thus organized a respective challenge based on our previously published dataset and baseline approaches (Hagen et al., 2022). With approaches submitted by 30 teams, the challenge had a very good participation. To improve reproducibility and replicability, we asked the teams to submit their software via the TIRA platform (Fröbe et al., 2023)—already used in several shared task with an archive of more than 500 research prototypes. Still, for our Clickbait Challenge, we also allowed result submissions, but a majority of the teams (19 of the 30) actually submitted software in the form of Docker images.

The task of clickbait spoiling, as exemplified in Figure 1, specifically aims at completing the "life cycle" of clickbait posts. The creation (Xu et al., 2019) and detection (Potthast et al., 2016, 2018) of clickbait posts are already operationalized, so we now aim to cover the "end" by automating spoiling. If sufficiently many people adopt spoiling tools, a clear message is send to publishers and social media platforms. Still, spoiling clickbait instead of removing it gives publishers the benefit of the doubt, since there may also be people who enjoy clicking on links for these kinds of trivia.



Clickbait tweet — **Spoiler**

Lifehacker @lifehacker
How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ

"washing [them]"

New York Post @nypost
Just how safe are NYC's water fountains? nyp.st/2yHSGnr

"The Post independently tested eight water fountains in New York City's most frequented parks, and found that all met or exceeded the state's guidelines for water quality."

CNBC @CNBC
A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus." (via @CNBCMakeIt) cnb.cx/2TG6zeX

"1. Added sugar" [...]
"2. Fried foods" [...]
"3. High-glycemic-load carbohydrates" [...]
"4. Alcohol" [...]
"5. Nitrates" [...]

Figure 1: Clickbait tweets and spoilers from the linked web pages (phrase, passage, and multipart spoiler).

The PAN Clickbait Challenge at SemEval 2023 has two subtasks: (1) spoiler type classification, and (2) spoiler generation. For spoiler type classification (i.e., should the spoiler be a phrase, a passage, or have multiple parts), the input consists of a clickbait post and the linked document. For spoiler generation, the input additionally includes the desired type of the spoiler. To motivate the teams to explore diverse approaches, the number of submissions of a team for either task was not limited. For spoiler generation, we focused on the evaluation of extractive spoilers (i.e., spoilers in the form of text from the linked document) and leave abstractive spoilers for future work (as conducting robust evaluations using generated texts would substantially increase the labeling efforts).

The datasets, results, and submitted software of the Clickbait Challenge 2023 are freely available.[2]

[1] Saved You a Click has 1.9 million members on Reddit.

[2] github.com/pan-webis-de/SEMEVAL-2023

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2275–2286
July 13-14, 2023 ©2023 Association for Computational Linguistics

2 Lab Overview and Statistics

For the clickbait spoiling task, 83 teams from 24 countries registered. The majority of registrations came from the United States (15 registered teams), followed by Germany and India (13 teams each), Bangladesh (8 teams), Romania (6 teams), China (4 teams), Poland (3 teams), Denmark, Israel, Mexico, Russia (2 teams each), Bulgaria, Canada, Colombia, France, Japan, Mexico, Morocco, Netherlands, Norway, Poland, Spain, Thailand, Turkey (1 team each). Overall, 26 of the 83 teams were undergrad students supervised by a senior researcher—most undergrad teams came from Germany (8)and India (7).

Of the 83 registered teams, 30 actively participated in the task, out of which 23 teams submitted notebook papers describing their approaches. We used TIRA (Fröbe et al., 2023) as our submission platform through which participants could either make software submissions or result submissions. Software submissions have advantages over result submissions in terms of reproducibility (i.e., software can be rerun on different data) but also allow for blinded experiments[3] (i.e., ensuring that the test set is kept private). While we allowed result submissions, we strongly encouraged software submissions and made the test data only available upon request. Overall, 10 active teams submitted results only, 1 team submitted results and software, while a majority of 19 teams submitted their software and never had access to the test data.

To submit a software, a team implemented their approach in a Docker image that they then uploaded to their dedicated Docker registry in TIRA. Software submissions in TIRA are immutable, and after uploading the image, the teams specified the to-be-executed command—the same Docker image can thus be used for multiple software submissions (e.g., by changing some parameters). A team could upload as many Docker images or software submissions as they liked; only they and TIRA had access to their dedicated Docker image registry (i.e., the images were not public while the shared task was ongoing). Overall, we received 202 software submissions from 86 unique Docker images (the smallest image has a compressed size of 504 MB, the largest image has 45 GB).

The teams could execute their software on the validation and the test data. On the validation data, the results, the evaluation scores, and the standard and error output were directly visible to be able to verify that a submitted software works as expected. On the test data, the evaluation scores and the outputs were blinded; the only available feedback was whether the software produced a valid output or not. To ensure that the test data is not leaked and to improve the overall reproducibility, TIRA executes software in a sandbox by removing the internet connection. This, for instance, helps to ensure that a software is fully installed in the Docker image which eases rerunning a software at some later point in time (all libraries and models must be contained in an image to be able to execute the software without internet connection). We assisted the teams in dockerizing their approaches (frequent issues were files that were not executable in an image, or resources like libraries or models that were not included in the image). Eventually, all teams who wanted to make a software submission actually submitted software that produced valid outputs. For the execution, participants could select resources out of four options: (1) 1 CPU core with 10 GB RAM, (2) 2 cores with 20 GB RAM, (3) 4 cores with 40 GB RAM, or (4) 1 CPU core with 10 GB RAM and 1 Nvidia GeForce GTX 1080 GPU with 8 GB RAM. Upon request, TIRA could even execute a software on a Kubernetes cluster with 1,620 CPU cores, 25.4 TB RAM, and 24 GeForce GTX 1080 GPUs. A team could run their software as often as they wanted using different resources to study the scalability and reproducibility (e.g., does a software run on a GPU yield the same results as on a CPU?).

3 Task Description and Datasets

In the PAN Clickbait Challenge at SemEval 2023, we follow our previous observation (Hagen et al., 2022) and distinguish three different types of spoilers: (1) short phrase spoilers, (2) longer passage spoilers, and (3) multi-part spoilers. Examples are given in Figure 1. Based on the hypothesis that tailored spoiler generators for the different spoiler types will be the most effective (i.e., identifying the required spoiler type before actually generating a spoiler might be beneficial), we organized two subtasks: (1) spoiler type classification, and (2) spoiler generation with given spoiler type information.

The input for spoiler type classification is the clickbait post and the linked document. The task is to identify whether the clickbait post requires either a phrase (the upper post in Figure 1), a passage (the

[3] wikipedia.org/wiki/Blinded_experiment

Table 1: Examples from Figure 1 as they would appear in our corpus (JSONL format in tabular form for readability).

UUID	**Clickbait Post**		**Linked Web Page**		**Spoiler**	
	Platform	Text	Title	Paragraphs	Type	Position
08. . .	Twitter	How to keep your workout clothes from stinking	How to Keep Your . . .	["Sweaty clothes stink, but . . . ", . . . , ". . . consider washing your stuff . . . "]	Phrase	[[[7, 276], [7, 283]]]
15. . .	Twitter	Just how safe are NYC's water fountains?	Just how safe are . . .	["The Post independently tested . . . ", . . . , "Still worried? For a cleaner . . . "]	Passage	[[[0, 0], [0, 171]]]
42. . .	Twitter	A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus."	A Harvard nutritionist and brain expert says . . .	["No matter how old you are . . . ", . . . , "1. Added sugar, as the brain . . . ", . . . , "2. Fried foods like French . . . ", . . . , "3. High-glycemic-load carb . . . ", . . .]	Multi	[[[3, 0], [3, 14]], [[6,0] [6,14]], [[10,0] [10,35]], . . .]

middle post in Figure 1), or a multi-part spoiler (the bottom post in Figure 1). For an input post (with a UUID), an output in the form of

```
{"uuid":"<UUID>", "spoilerType":"<TYPE>"}
```

has to be generated, where `<TYPE>` is either `phrase`, `passage`, or `multi`.

The input for spoiler generation is the clickbait post, the linked document, and the required spoiler type (`phrase`, `passage`, or `multi`). For an input post (with a UUID), an output in the form of

```
{"uuid":"<UUID>", "spoiler":"<SPOILER>"}
```

has to be generated, where `<SPOILER>` is the spoiler for the clickbait post.

Webis Clickbait Spoiling Corpus 2022 As the data for the tasks, we use our Webis Clickbait Spoiling Corpus 2022 (5,000 posts with spoilers, fixed random 64/16/20 train/validation/test split for both subtasks). The training and validation sets were publicly available while the test dataset was made available upon request only. After the second Clickbait Challenge has ended, the complete corpus will be accessible under a research-friendly Creative Commons Attribution license in JSONL format.[4]

Table 1 showcases the main corpus fields for the posts from Figure 1. Since no main content extraction method worked reliably for all the linked web pages, we manually extracted titles and paragraphs while annotating the spoiler positions and types. In total, we have spent about 560 hours of in-house work on manual post selection, main content extraction, and spoiler identification to ensure high data quality (details on the corpus: (Hagen et al., 2022)). Most spoilers are phrases (42.5%) or passages (40%)—our annotation guidelines asked for a spoiler to be as short as possible (i.e., if one word is enough, not a whole sentence should be chosen). A spoiler's exact position in the linked page is part of the corpus, so that an automatic assessment of extractive spoilers using BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020) is possible.

4 Task 1: Spoiler Type Classification

The goal of the first task of the PAN Clickbait Challenge at SemEval 2023 was to classify which spoiler type a given clickbait post requires. We explored a wide range of approaches during a pilot study to provide a competitive baseline as the basis for the shared task. We evaluated the effectiveness of spoiler type classification using the balanced accuracy over all three classes as the main effectiveness measure that we complement by the precision, recall, and F1 score in detecting all three classes. Overall, 25 teams submitted results to task 1.

4.1 Baseline and Pilot Experiments

For spoiler classification, we experimented with the classic feature-based models Naïve Bayes, Logistic Regression, and SVM, as well as the neural models BERT (Devlin et al., 2019), DeBERTa (He et al., 2021), and RoBERTa (Liu et al., 2019). The classic models used tf- and $tf \cdot idf$-weighted word and POS tag uni- and bigrams from the clickbait post and the linked web page (chi-square feature selection). For the neural models, we concatenated the post and the main content of the linked page to predict the spoiler type. Logistic regression as the best classic model was less effective than the best neural model RoBERTa (61.10 vs. 73.40 balanced accuracy on the validation dataset). Hence, we used RoBERTa as the baseline for task 1, as it was the most effective model but still leaves much room for further improvements. We released the code and the model of the baseline for task 1 as well as the resulting Docker image to simplify the

[4] doi.org/10.5281/zenodo.6362726

onboarding for participants.[5]

4.2 Participating Systems

Team Alexander Knox (Woźny and Lango, 2023) explored multiple prompts, and augmented spoilers for classification approaches using few-shot learning with large language models or finetuning of transformers. The large language model BLOOM (Scao et al., 2022) was used to augment a hypothetical spoiler for a given clickbait post and article, where BLOOM generated a phrase spoiler, a passage spoiler, and a multi-part spoiler in a few-shot setting. The generated spoilers were then used as additional input for the spoiler classification task, and pilot experiments showed that few-shot classification with BLOOM was substantially less effective than finetuned transformers DistilBERT-base (Sanh et al., 2019), ALBERT, and RoBERTa-base were used where RoBERTa achieved the highest effectiveness.

Billie Newman (Kruff and Tran, 2023) used RoBERTa and additional features extracted by specifically designed rules. Regular expressions were used to extract the number of enumerations, lists, named entities, and a ratio that compares the length of the clickbait tweet to the length of the linked article, which was then added as additional textual context to the clickbait post and linked article before they were passed to the RoBERTa model.

Team Billy Batson (Sharma et al., 2023) used the clickbait post and the linked article as input to a DeBERTa-base to classify the spoiler type where paragraphs were ranked by their relevance to the clickbait post using a pairwise retrieval model. The retrieval model scored all passages of the linked article using a RoBERTa model that was trained in a contrastive learning setup (telling a passage containing a spoiler apart from a passage without a spoiler), retaining only the top-k passages as input for the DeBERTa based type classification.

Team Chick Adams (Pan et al., 2023) used different inputs to RoBERTa and DeBERTa models. Using either only the clickbait post, the linked article, or both as input to either RoBERTa respectively DeBERTa, they found that DeBERTa was more effective than RoBERTa and that the concatenation of the clickbait post and the linked article were more effective then the clickbait post or the linked article alone, and also that the clickbait post and the linked article do not complement each other as ensembles of both were less effective in their pilot experiments. The most effective model was DeBERTa using the concatenation of the clickbait post and the linked article as input.

Team Clark Kent (Mihalcea and Nisioi, 2023) experimented with the Pixel-based Encoder of Language (PIXEL) model (Rust et al., 2022) for task 1 to inspect if visual properties of clickbait posts and articles help in spoiler classification. PIXEL transforms its input text into RGB images and is a pre-trained vision transformer masked autoencoder with 112M parameters. The PIXEL model was finetuned for multiple epochs on the training data with early stopping using the clickbait post and the linked article as input.

Team Mr. Fosdick (Falkenberg et al., 2023) experimented with data augmentation and expansion techniques, deriving new datasets from the original training dataset by paraphrasing the clibkbait posts and the clickbait articles with GPT3. The augmented datasets where then used to train a transformer model and a random forrest model, where a DeBERTa model that used the clcikbait post and the linked article as input was the most effective approach trained on the augmented dataset.

Team Francis Wilde (Indurthi and Varma, 2023) used RoBERTa for spoiler-type classification in task 1. The RoBERTa model used the clickbait post as input (the linked article is ommitted) and was finetuned on the official training dataset.

Team Gallagher (Bilgis et al., 2023) experimented with T5, Long-T5, and Flan-T5, finding that a T5-Large model was more effective than Long-T5 in classifying the spoiler type by predicting the probability that the clickbait post and linked article as input generate the token phrase, passage, respectively multi.

Jack Flood (Kumar et al., 2023) used a BiLSTM model that used the length of the clickbait post and the linked article, the count of overlapping terms between the post and the article, and RoBERTa representations of the post and article as input.

Team John Boy Walton (Shmalts, 2023) used an ensemble of five models (variants of DistilBERT, BERT-base, and DeBERTa), averaging the logits of the five models to obtain the final classification.

Team Machamp trained a multi-task-learning model using the Machamp framework (van der Goot et al., 2021) on all SemEval-2023 tasks following the idea that training a single model on di-

[5] github.com/pan-webis-de/pan-code/tree/master/semeval23/baselines

verse tasks enables synergies between the different training tasks.

Team Matt Bai (Tailor and Mamidi, 2023) used BERT-base to classify the spoiler type using the clickbait post and linked article as input to the model (a second submission used only the clickbait post as input but this was less effective).

Team Mr. Wallace (Saravanan and Wilson, 2023) used a DistilBERT-base model using the clickbait post and linked article as input.

Team Monique Marmelstein (Sterz et al., 2023) used RoBERTa in a multi-task learning setup to train the model in parallel on task 1 and task 2 following the hypothesis that improvements in one task transfer to the other task as well.

Team Morbo the Annihilator experimented with Naíve Bayes, Logistic Regression, and Support Vector Machines comparing them under different input representations and preprocessing steps. The best approach used logistic regression using the term frequency and TF-IDF on the targetTitle, targetParagraphs, postText, topic modeling using matrix factorization on the concatenation of "targetTitle" and "postText", calculating similarity between targetTitle and postText, counting nouns and punctuation signs from targetTitle and.

Team Nancy Hicks Gribble (Keller et al., 2023) used the clickbait post and the linked article as input to RoBERTa models, comparing a multiclass classification model with one-vs-rest models, finding that the single multiclass RoBERTa model was overall more effective then combinations of three standalone models.

Team Perry White tested BERT, RoBERTa, and DeBERTa for spoiler type classification, exploring three types of data presentations: (1) using the article's contents only, (2) concatenating the beginning of the post, title of article and the articles contents, and truncating the rest, (3) were using posts text and only 200 first tokens from the article, skipping the title and first paragraph. Despite this method producing inputs shorter than the other two methods, it scores several p.p. higher. (this is inspired by the F-shape reading pattern theory[6]). The DeBERTa model was the most effective approach for the task (Post text + 200 first tokens yielded 76% accuracy after 3 epochs).

Team Sam Miller (Störmer et al., 2023) fine-tuned XLNET (Yang et al., 2019) using the clickbait post as input for spoiler classification.

Team Stephen Colbert (Spreitzer and Tran, 2023) used the newest version of DeBERTa-Large (version 3) while exploring different representations of the input dataset. The main idea was to use markup information from the linked article as additional input following the hypothesis that different types of clickbait posts use different techniques for search engine optimization that are manifested in the underlying markup of the linked article.

Team Walter Burns (Villa Cueva et al., 2023) uses the clickbait post and the title of the linked article as input to to an ensemble of 5 RoBERTa models that were trained from different random initializations. The class probabilities of the 5 RoBERTa models are then used as input for a neural network layer that predicts the spoiler type.

Team Frankly Unctuous, Miles Clarks, Paul Morgan, Spider Jerusalem, and Trinity Wells did not submit a notebook paper. We are in contact with the team to get a accurate high-level description of their approach. (Miles Clarks, Spider Jerusalem, and Trinity Wells made software submissions so that their submission is reproducible and could be inspected.)

4.3 Results

Table 2 compares each team's most effective submission for spoiler type prediction measured as balanced accuracy (the main measure that we employ; also the most effective submission for each team was selected via balanced accuracy) over all three spoiler types and the precision, recall, and F1 score for phrase, passage, and multi spoilers on the test set. Our baseline was very effective, and only team Billy Batson outperformed the baseline in the balanced accuracy but only in the third decimal place (that we do not report in the table to not overstate the precision of observations). We leave out all submissions that resubmitted the baseline to only report the distinct set of submissions. While the overall best accuracy of 0.74 shows that there is still much performance, inspecting the precision and recall of different approaches for the different spoiler types shows that the different approaches might complement each other well, which might be suitable directions for future work. For instance, the approach by Billy Batson achieves very high precision for detecting phrase and passage spoilers, while only achieving a precision of 0.58 for multi spoilers, while other approaches achieve substan-

[6] www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

Table 2: Overview of the effectiveness in spoiler type prediction (subtask 1 at SemEval 2023 Task 5) measured as balanced accuracy over all three spoiler types and precision (Pr.), recall (Rec.), and F1 score (F1) for phrase, passage, and multi spoilers on the test set. For each team, we only report the most effective submission.

Submission			**Accuracy**	**Phrase**			**Passage**			**Multi**		
Team	Name	Run		Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Billy Batson	result upload	02-13-07-44-28	0.74	0.79	0.70	0.74	0.76	0.74	0.75	0.58	0.79	0.67
Baseline	citron-landaulet	11-20-15-48-19	0.74	0.76	0.75	0.76	0.73	0.76	0.74	0.74	0.70	0.72
Gallegher	task1-gpu	02-01-23-07-17	0.72	0.73	0.77	0.75	0.76	0.69	0.72	0.66	0.70	0.68
Paul Morgan	result upload	12-17-16-10-28	0.72	0.75	0.74	0.75	0.72	0.77	0.74	0.74	0.63	0.68
Walter Burns	result upload	01-26-00-53-14	0.72	0.75	0.79	0.77	0.75	0.69	0.72	0.65	0.67	0.66
Perry White	deberta-t2	01-12-23-53-20	0.71	0.77	0.72	0.74	0.70	0.81	0.75	0.73	0.60	0.66
Spider Jerusalem	latent-tempo	02-01-21-19-12	0.71	0.76	0.74	0.75	0.69	0.81	0.75	0.88	0.57	0.69
Chick Adams	result upload	01-23-17-24-54	0.71	0.77	0.72	0.74	0.71	0.79	0.75	0.69	0.60	0.64
Alexander Knox	cyan-winner	01-24-09-50-19	0.70	0.71	0.78	0.75	0.73	0.72	0.72	0.78	0.60	0.68
Stephen Colbert	flat-cafe	02-03-18-57-38	0.70	0.75	0.74	0.74	0.71	0.79	0.75	0.76	0.57	0.65
Francis Wilde	result upload	01-25-04-55-22	0.70	0.74	0.69	0.71	0.68	0.78	0.73	0.74	0.64	0.69
Nancy Gribble	Inj-task1	01-24-08-06-24	0.70	0.78	0.66	0.71	0.67	0.84	0.75	0.75	0.59	0.66
Mr. Fosdick	Deberta	01-11-21-00-59	0.69	0.73	0.70	0.71	0.73	0.67	0.70	0.54	0.71	0.61
M. Marmelstein	chilly-parakeet	01-24-22-18-23	0.66	0.66	0.81	0.73	0.74	0.64	0.68	0.70	0.55	0.61
Sam Miller	bright-bicycle	01-31-07-46-19	0.66	0.78	0.57	0.66	0.65	0.80	0.72	0.57	0.62	0.59
Machamp	result upload	01-27-22-44-33	0.65	0.69	0.72	0.70	0.66	0.72	0.69	0.76	0.52	0.62
Miles Clarkson	test	01-14-01-13-55	0.65	0.77	0.60	0.68	0.72	0.60	0.65	0.40	0.76	0.52
Clark Kent	grouchy-matrix	02-09-21-28-31	0.65	0.69	0.71	0.70	0.67	0.73	0.70	0.70	0.51	0.59
Matt Bai	result upload	01-30-17-26-03	0.63	0.63	0.77	0.69	0.67	0.64	0.65	0.85	0.47	0.61
Billie Newman	s.-chutney	01-25-20-17-47	0.59	0.78	0.30	0.43	0.60	0.83	0.70	0.40	0.65	0.50
John Walton	legato-stress	02-28-01-27-05	0.58	0.59	0.72	0.64	0.62	0.62	0.62	0.85	0.40	0.54
Morbo	desert-style	01-15-09-51-48	0.54	0.56	0.70	0.63	0.61	0.58	0.59	0.66	0.33	0.44
Jack Flood	result upload	02-12-16-10-51	0.52	0.57	0.56	0.56	0.55	0.66	0.60	0.61	0.34	0.44
Mr. Wallace	result upload	01-23-01-10-16	0.50	0.52	0.65	0.58	0.55	0.57	0.56	0.78	0.29	0.43
Trinity Wells	colorful-vertex	01-15-09-49-12	0.33	0.00	0.00	0.00	0.40	1.00	0.57	0.00	0.00	0.00
F. Unctuous	abs.-recursion	01-26-21-11-33	0.33	0.00	0.00	0.00	0.40	1.00	0.57	0.00	0.00	0.00

tially higher precision for multipart spoilers.

Furthermore, the precision and recall scores per spoiler type in Table 2 are particularly important to guide real implementations of spoiling algorithms. First practical implementations might focus specifically on precision to ensure that wrong spoilers do not increase users efforts, e.g., by only spoiling clickbait posts requiring a certain type of easy-to-spoil clickbait posts.

5 Task 2: Spoiler Generation

The goal of the second task of the PAN Clickbait Challenge at SemEval 2023 was to generate the actual spoiler that could be shown to users to satisfy the curiosity introduced by a clickbait post. We explored a wide range of approaches during a pilot study to provide a competitive baseline as the basis for the shared task. We evaluated the effectiveness of spoiler generation by contrasting BLEU-4, METEOR, and BERTScore scores (generated vs. ground-truth spoiler). Overall, 17 teams submitted results to task 2.

5.1 Baseline and Pilot Experiments

For spoiler generation, we experimented with the question answering approaches BERT, DeBERTa, and RoBERTa, as well as the passage retrieval approaches BM25, MonoBERT (Nogueira and Cho, 2019; Nogueira et al., 2019), and MonoT5 (Nogueira et al., 2020). The question answering approaches used the clickbait post as question aiming to find the spoiler as answer in the linked article. We found that question answering approaches are substantially more effective than passage retrieval approaches and that fine-tuning the question answering models on SQuAD before fine-tuning on our clickbait spoiling task improves their effectiveness. We also compared models simply trained on all training posts to two-step approaches that first classify the spoiler type to then select a spoiler generator only trained on the respective spoiler type. For the two-step approaches, we compare variants using a perfect "oracle" classification or the RoBERTa classifier that achieved the best accuracy in the classification pilot study. Overall, DeBERTa achieves the highest spoiling effectiveness (according to BLEU-4, METEOR, and BERTScore) in all scenarios. We released the code and the model of the baseline for task 2 as well as the resulting Docker image to simplify the onboarding for participants.[7]

5.2 Participating Systems

Team Billie Newman (Kruff and Tran, 2023) participated used a mixture of RoBERTa models enriched with a rule-based approach following manually designed rules. Two RoBERTa models were first fine-tuned on SQuAD version 2 for extractive question answering and subsequently one model was finetuned for phrase spoiling, and one model for passage spoiling, while a rule-based approach was used to spoil multi-type spoilers.

Team Brooke English (Tang, 2023) used four DeBERTa models for spoiler generation that were selected using the given spoiler type and the confidence score of the DeBERTa model. Three dedicated DeBERTa models were trained for phrase, passage, respectively multi-part spoilers, and one DeBERTa model for all three types. Each model used the clickbait post concatenated to the linked article as input, and for each input, two spoiler candidates were generated, one with the dedicated DeBERTa model for the spoiler type and one with the general-purpose DeBERTa model, and the candidate with the higher probability as estimated by the generating model was returned as spoiler.

Team Diane Simmons (Krog and Agirrezabal, 2023) experimented with DeBERTa in a finetuning setup and the large language model BLOOM in a zero-shot setup. DeBERTa used the clickbait post and the linked article as input to extract the spoiler from the linked article, while the clickbait post and the linked article where embedded in diverse prompts for abstractive spoiler generation using BLOOM. The DeBERTa model was substantially more effective then BLOOM.

Team Gallagher (Bilgis et al., 2023) experimented with T5 (Raffel et al., 2020), Long-T5 (Guo et al.), and Flan-T5 (Chung et al., 2022) building an ensemble: T5-Large was used for phrase and multi spoilers (was more effective then Long-T5 and Flan-T5) while Flan-T5-Large was used for passage spoilers.

Team Jack Flood (Kumar et al., 2023) used passage retrieval appraoches to select relevant passages from the linked article that are subsequently used as input to a RoBERTa model for spoiling. The linked article is segmented into sentences that are ranked with reciprocal rank fusion combining monoT5 and BM25 scores. The top-5 sentences are subsequently passed to dedicated RoBERTa models for phrase, passage, respectively multipart spoilers.

Team Jack Ryder (Wangsadirdja et al., 2023) use multiple zero-shot question-answering models without fine-tuning as ensemble, reformulating the clickbait posts as questions to imitate the expected input of pre-trained question-answering models showing the sentences from the linked article sorted by their SBERT similarity. They experiment with RoBERTa, DeBERTa, Flan-T5, and UnifiedQA, finding in pilot studies that rephrasing the clickbait post as questions usually improves the effectiveness while sorting sentences by their SBERT similarity does not. The final model combines the best configuration for each spoiler type used in a zero-shot setup.

Team John Boy Walton (Shmalts, 2023) ensembled five models (variants of DistilBERT, BERT-base, and DeBERTa), averaging the logits of the five models to obtain the extractive spoiler (extracting the position of the start and end of the spoiler).

Team John King (Kurita et al., 2023) used an ensemble of five sequence-to-sequence models (T5-base, T5-large, Flan-T5-base, flan-t5-large, and DeBERTa-large) for clickbait spoiling following the idea that those models can generate multi-part spoilers without adaption (e.g., span detection approaches would require more complex adoptions for multi-part spoilers). Each sequence-to-sequence model generated spoiler candidates independent of each other, returning the candidate with the minimum edit distance to all other candidates as the spoiler.

Team Morbo the Annihilator used RoBERTa-base for spoiling, starting from Roberta-base finetuned on SQUAD version 2. Three models were trained, for phrase, passage, and multi-part spoilers that used the clickbait post concatenated to the linked article as input.

Team Monique Marmelstein (Sterz et al., 2023) used RoBERTa in a multi-task learning setup to train the model in parallel on task 1 and task 2 following the hypothesis that improvements in one task transfer to the other task as well.

Team Nancy Hicks Gribble (Keller et al., 2023) used a RoBERTa model that was pre-trained on

[7] github.com/pan-webis-de/pan-code/tree/master/semeval23/baselines

the SQuAD (Rajpurkar et al., 2018) dataset for question answering (roberta-base-squad2[8]). Three different models were trained for each spoiler type, using the clickbait post and the linked article as input.

Team Sabrina Spellman (Birkenheuer et al., 2023) trained a DeBERTa-base model using an additional dataset constructed from Huffington Post articles to better learn the structure of news articles. The first distant supervision dataset for training was inspired by cloze-style questions, in that a passage was removed from an article to fill a gap in a given sentence. The second distant supervision dataset for training retrieved enumerations from the text. After fine-tuning DeBERTa on both distant-supervision tasks, three models were trained specifically for phrase, passage, and multi spoilers.

Team Walter Burns (Villa Cueva et al., 2023) used an ensemble of RoBERTa and DeBERTa by averaging the predicted logits of both models. For multi-part spoiling, the top-5 extractive, non-overlapping spoilers were extracted from the averaged logits.

5.3 Results

Table 3 compares each teams most effective submission for spoiler generation measured as BLEU-4 (BL4), BERTScore (BSc.), and METEOR (MET) accross all clickbait posts and across the three spoiler types. In contrast to task 1, much more teams outperform our strong baseline (6 teams have submissions with a higher BLEU-4 score). Team John King achieves the highest overall effectiveness (BLEU-4 of 0.48). Remarkably, the best approach of John King uses a generative sequence-to-sequence approach, although our evaluation should favor extractive spoiling approaches as we calculate the scores in comparison to the ground-truth spoiler extracted from the linked article. Comparing the effectiveness accross the different spoiler types reveals that there is still substantial room for further improvement: While John King clearly achieves the highest effectiveness for multi spoilers, other submissions by Sabrina Spellman and Walter Burns achieve comparable effectiveness for phrase spoilers and the approach by Sabrina Spellmann substantially outperforms the submission of John King for passage spoilers. Interestingly, the intention of team John King to use a sequence-to-sequence model for multipart spoiling because they dont need complex adoptions in comparison to span based extractive approaches proved to be correct because those models are substantially more effective then all other models (e.g., team John King achieves a BLEU-4 score of 0.44 that is followed 0.30 of team Sabrina Spellman).

```
tira.run(
    'clickbait-spoiling/<team-name>/<software>',
    dataset='<dataset>'
)
```

Listing 1: Local re-execution of the software `<software>` by team `<team-name>` that was submitted to the clickbait spoiling task.

Combining the results from Table 2 and Table 3 reveals that first user-facing implementations of clickbait spoiling might focus on only spoiling phrase spoilers, as the precision oriented detection of phrase spoilers is possible (maximum precision of 0.79) and spoiling phrase spoilers is also very accurate (the approach of team gallagher achieves an BLEU-4 score of 0.69, a BERTScore of 0.96, and a METEOR score of 0.71).

6 Post-Hoc Reproducibility Experiments

We used TIRA (Fröbe et al., 2023) as submission system so that we can publish all artifacts resulting from the shared task to improve reproducibility and re-usability. During the shared task, TIRA maintained all data in a Git repository, where each software execution and evaluation were triggered by commits in continuous integration and continuous deployment pipelines. We publish this repository, including all datasets, results, evaluations, logs, metadata, and software snapshots.[9]

Listing 1 exemplifies how software submissions can be re-executed on the same or new data (the API allows the re-use of existing datasets but also new data of the same structure, requiring only Docker and Python 3 as dependencies). Consequently, the shared task repository that we published after the completion of the shared task now serves as a entry point for follow-up studies. All researchers can fork this repository and contribute.

7 Conclusion

The second PAN Clickbait Challenge hosted as Task 5 at SemEval 2023 received submission from 30 active teams on two subtasks: (1) spoiler type

[8] huggingface.co/deepset/roberta-base-squad2

[9] github.com/pan-webis-de/SEMEVAL-2023

Table 3: Overview of the effectiveness in spoiler generation (subtask 2 at SemEval 2023 Task 5) measured as BLEU-4 (BL4), BERTScore (BSc.) and METEOR (MET) over all clickbait posts respectively those requiring phrase, passage, or multi spoilers on the test set. For each team, we only report the most effective submission.

Submission			**All**			**Phrase**			**Passage**			**Multi**		
Team	Name	Run	BL4	BSc.	MET	BL4	BSc.	MET	BL4	BSc.	MET	BL4	BSc.	MET
John King	burning-desert	02-06-07-29-13	0.48	0.93	0.50	0.66	0.95	0.40	0.32	0.91	0.43	0.44	0.93	0.65
S. Spellman	cerulean-r.	01-24-00-32-15	0.47	0.93	0.50	0.64	0.95	0.69	0.36	0.91	0.51	0.30	0.90	0.45
Walter Burns	result upload	01-26-01-01-19	0.44	0.92	0.49	0.68	0.96	0.58	0.28	0.90	0.47	0.21	0.90	0.48
Mr. Fosdick	Squad QA	01-10-21-21-07	0.42	0.92	0.48	0.65	0.95	0.46	0.32	0.91	0.50	0.12	0.89	0.43
Brooke English	congruent-eel	01-31-05-27-54	0.42	0.92	0.42	0.66	0.95	0.41	0.29	0.90	0.44	0.15	0.89	0.39
Gallegher	task2-gpu	02-01-21-35-34	0.41	0.92	0.44	0.69	0.96	0.71	0.24	0.90	0.42	0.12	0.88	0.35
Baseline	legato-cinema	01-01-10-57-35	0.40	0.92	0.43	0.65	0.95	0.60	0.24	0.90	0.44	0.12	0.87	0.30
M. Marmelstein	beige-cordon	01-31-23-01-02	0.36	0.91	0.43	0.62	0.95	0.55	0.19	0.89	0.41	0.11	0.89	0.38
Morbo	achromatic-f.	02-01-08-16-39	0.36	0.91	0.39	0.63	0.95	0.65	0.19	0.89	0.36	0.09	0.88	0.30
Billie Newman	poky-sprite	01-24-17-37-56	0.32	0.90	0.30	0.49	0.92	0.23	0.20	0.87	0.31	0.20	0.88	0.34
Sam Miller	equidistant-m.	02-01-04-47-19	0.31	0.66	0.38	0.60	0.95	0.54	0.11	0.55	0.27	0.10	0.22	0.46
Jack Ryder	joint-trap	02-01-12-59-21	0.27	0.89	0.27	0.48	0.92	0.37	0.16	0.88	0.28	0.03	0.85	0.18
N. Gribble	short-screw	01-28-17-53-10	0.27	0.88	0.26	0.48	0.93	0.44	0.14	0.84	0.25	0.05	0.85	0.23
Diane Simmons	result upload	01-24-14-49-38	0.25	0.89	0.26	0.48	0.93	0.44	0.07	0.86	0.23	0.08	0.85	0.18
Jack Flood	result upload	02-12-16-12-01	0.18	0.88	0.18	0.32	0.89	0.14	0.08	0.87	0.21	0.05	0.85	0.16
J. Walton	felt	02-13-16-20-23	0.08	0.86	0.14	0.16	0.87	0.17	0.03	0.85	0.14	0.01	0.83	0.10
Miles Clarkson	black-cadet	01-24-09-32-18	0.04	0.84	0.10	0.04	0.84	0.04	0.02	0.84	0.12	0.07	0.85	0.17
Machamp	result upload	01-28-15-24-37	0.00	0.83	0.01	0.00	0.85	0.00	0.00	0.82	0.01	0.00	0.81	0.01

classification to assess what kind of spoiler a clickbait post requires, and (2) spoiler generation to generate an actual spoiler for a clickbait post. The most effective spoiler type classifier uses sentence retrieval to run a DeBERTa-based classifier on the most relevant sentences of the linked document. The most effective spoiler generator uses five independently run generative sequence-to-sequence spoiler generators, from whose outputs the most promising spoiler is selected via majority voting based on edit distances.

Our results show some substantial advancements, but there is still much room for further improvements. First, as no spoiler generator achieves the highest effectiveness across all three spoiler types, the best generators still complement each other and combining the underlying ideas can probably further increase the effectiveness. Second, as the most effective spoiler generator is ensemble-based and comes as a 45 GB Docker image, distilling a model with a better efficiency–effectiveness tradeoff also is an interesting direction for future work.

Acknowledgements

We want to thank all participants of the shared task for their valuable submissions. Furthermore, we thank all reviewers: Alexander Bondarenko, Anubhav Sharma, Dirk Wangsadirdja, Dragos Stefan Mihalcea, Ferdinand Schlatt, Hiroto Kurita, Ikumi Ito, Jüri Keller, Kim Bürgl, Maksim Shmalts, Manex Aguirrezabal Zabaleta, Mateusz Woźny, Sebastian Günther, Sujit Kumar, Vijayasaradhi Indurthi, and Vineet Saravanan.

This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Tugay Bilgis, Nimet Beyza Bozdag, and Steven Bethard. 2023. Gallagher at semeval-2023 task 5: Tackling clickbait with seq2seq models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1650–1655, Toronto, Canada. Association for Computational Linguistics.

Simon Birkenheuer, Jonathan Drechsel, Paul Justen, Jimmy Pöhlmann, Julius Gonsior, and Anja Reusch. 2023. Sabrina spellman at semeval-2023 task 5: Discover the shocking truth behind this composite approach to clickbait spoiling! In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 969–977, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. 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. *CoRR*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Christian Falkenberg, Erik Schönwälder, Tom Rietzke, Chris-Andris Görner, Robert Walther, Julius Gonsior, and Anja Reusch. 2023. Mr-fosdick at semeval-2023 task 5: Comparing dataset expansion techniques for non-transformer and transformer models: Improving model performance through data augmentation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 88–93, Toronto, Canada. Association for Computational Linguistics.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*.

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)*. Association for Computational Linguistics. (to appear, preprint: `http://arxiv.org/pdf/2203.10282`).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Vijayasaradhi Indurthi and Vasudeva Varma. 2023. Francis wilde at semeval-2023 task 5: Clickbait spoiler type identification with transformers. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1890–1893, Toronto, Canada. Association for Computational Linguistics.

Jüri Keller, Nicolas Rehbach, and Ibrahim Zafar. 2023. nancy-hicks-gribble at semeval-2023 task 5: Classifying and generating clickbait spoilers with roberta. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1712–1717, Toronto, Canada. Association for Computational Linguistics.

Niels Krog and Manex Agirrezabal. 2023. Diane simmons at semeval-2023 task 5: Is it possible to make good clickbait spoilers using a zero-shot approach? check it out! In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 477–481, Toronto, Canada. Association for Computational Linguistics.

Andreas Kruff and Anh Huy Tran. 2023. Billie-newman at semeval-2023 task 5: Clickbait classification and question answering with pre-trained language models, named entity recognition and rule-based approaches. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1542–1550, Toronto, Canada. Association for Computational Linguistics.

Sujit Kumar, Aditya Sinha, Soumyadeep Jana, Rahul Mishra, and Sanasam Ranbir Singh. 2023. Jackflood at semeval-2023 task 5:hierarchical encoding and reciprocal rank fusion-based system for spoiler classification and generation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1906–1915, Toronto, Canada. Association for Computational Linguistics.

Hiroto Kurita, Ikumi Ito, Hiroaki Funayama, Shota Sasaki, Shoji Moriya, Ye Mengyu, Kazuma Kokuta, Ryujin Hatakeyama, Shusaku Sone, and Kentaro Inui. 2023. Tohokunlp at semeval-2023 task 5: Clickbait spoiling via simple seq2seq generation and ensembling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1756–1762, Toronto, Canada. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Dragos-Stefan Mihalcea and Sergiu Nisioi. 2023. Clark kent at semeval-2023 task 5: Svms, transformers, and pixels for clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1204–1212, Toronto, Canada. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR*, abs/1901.04085.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings* *of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. *CoRR*, abs/1910.14424.

Ronghao Pan, José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023. Chick adams at semeval-2023 task 5: Using roberta and deberta to extract post and document-based features for clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 624–628, Toronto, Canada. 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 of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. 2018. The Clickbait Challenge 2017: Towards a Regression Model for Clickbait Strength. *CoRR*, abs/1812.10847.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. In *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York. Springer.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2022. Language modelling with pixels. *CoRR*, abs/2207.06991.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Vineet Saravanan and Steven Wilson. 2023. Mr-wallace at semeval-2023 task 5: Novel clickbait spoiling algorithm using natural language processing. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1625–1629, Toronto, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Anubhav Sharma, Sagar Joshi, Tushar Abhishek, Radhika Mamidi, and Vasudeva Varma. 2023. Billy-batson at semeval-2023 task 5: An information condensation based system for clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1878–1889, Toronto, Canada. Association for Computational Linguistics.

Maksim Shmalts. 2023. John boy walton at semeval-2023 task 5: An ensemble approach to spoiler classification and retrieval for clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2100–2106, Toronto, Canada. Association for Computational Linguistics.

Sabrina Spreitzer and Hoai Nam Tran. 2023. Stephen colbert at semeval-2023 task 5: Using markup for classifying clickbait. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1844–1848, Toronto, Canada. Association for Computational Linguistics.

Hannah Sterz, Leonard Bongard, Tobias Werner, Clifton Poth, and Martin Hentschel. 2023. Ml mob at semeval-2023 task 5: "breaking news: Our semi-supervised and multi-task learning approach spoils clickbait". In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1818–1823, Toronto, Canada. Association for Computational Linguistics.

Pia Störmer, Tobias Esser, and Patrick Thomasius. 2023. Sam miller at semeval-2023 task 5: Classification and type-specific spoiler extraction using xlnet and other transformer models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1217–1224, Toronto, Canada. Association for Computational Linguistics.

Nukit Tailor and Radhika Mamidi. 2023. Matt bai at semeval-2023 task 5: Clickbait spoiler classification via bert. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1067–1068, Toronto, Canada. Association for Computational Linguistics.

Shirui Tang. 2023. Brooke-english at semeval-2023 task 5: Clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 64–76, Toronto, Canada. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Emilio Villa Cueva, Daniel Vallejo Aldana, Fernando Sánchez Vega, and Adrián Pastor López Monroy. 2023. Walter burns at semeval-2023 task 5: Nlp-cimat - leveraging model ensembles for clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 693–699, Toronto, Canada. Association for Computational Linguistics.

Dirk Wangsadirdja, Jan Pfister, Konstantin Kobs, and Andreas Hotho. 2023. Jack-ryder at semeval-2023 task 5: Zero-shot clickbait spoiling by rephrasing titles as questions. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1090–1095, Toronto, Canada. Association for Computational Linguistics.

Mateusz Woźny and Mateusz Lango. 2023. Alexander knox at semeval-2023 task 5: The comparison of prompting and standard fine-tuning techniques for selecting the type of spoiler needed to neutralize a clickbait. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1470–1475, Toronto, Canada. Association for Computational Linguistics.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3063–3073. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.