Clickbait Spoiler Classification and Spoiler Generation

Himani Madan
University at Buffalo
himani@buffalo.edu

Abstract

Clickbait headlines entice audiences to click on irrelevant information, resulting in the wastage of users' time. To address this issue, spoilers can be generated and presented to users, allowing them to consume precise information without going through a plethora of useless content. In this study, we propose a classification for clickbait spoilers, categorizing them into three types: phrase, passage, and multi. We also developed methods to generate or extract spoilers from target paragraphs within clickbait articles. Our experimental results using Roberta base [7] demonstrate a spoiler classification with F1 score of 70%. Additionally, employing T5 base [2], we achieve a bleu score exceeding 50% for phrase-type spoiler generation, and 89% for passage and 81% multi-type spoiler generation using T5 long model.

1 Introduction

The internet plays a pivotal role in information dissemination and communication, where social media platforms have become key connectors between users and online content. In this digital landscape, clickbait content has emerged as a distinct form of online content, strategically designed to capture attention and generate clicks. Clickbait often employs sensationalized headlines or thumbnails, utilizing misleading or exaggerated claims to entice users into clicking on the accompanying link. However, upon clicking, users often find that the actual content fails to live up to the initial expectations set by the clickbait. Consequently, there is a need to address this issue and prevent users from wasting their time on lengthy and uninformative content.

In our research, we delve into the exploration of clickbaits and the use of spoilers as a potential solution. To conduct our investigation, we utilize the Webis Clickbait Spoiling Corpus 2022 dataset. This dataset comprises a collection of 5,000 posts obtained from popular social media platforms such as Twitter, Reddit, and Facebook. These posts include both clickbaits and their corresponding spoiled clickbait content. The dataset offers a diverse and extensive range of clickbait topics, making it a valuable resource for researchers studying clickbait spoiling.

Out of the 5,000 posts in the dataset, we allocate 3,200 for training purposes, 800 for validation, and 1,000 for testing (unavailable to us during the research process). Initially, our focus revolves around the classification of spoilers into three distinct types: phrase, multi, and passage. Once the spoiler type is determined, the subsequent task involves generating spoilers that correspond to the associated clickbaits.

By undertaking this research, we aim to shed light on the effectiveness of spoilers in combating clickbait-related issues. Through our analysis of the dataset and subsequent experiments, we seek to gain insights into the classification and generation of spoilers, ultimately contributing to the development of effective strategies to tackle clickbait and enhance the online user experience.

2 Related Work

Traditional methods for clickbait detection have relied heavily on feature engineering which, in turn, is dependent on the dataset it is built for. The application of neural networks for this task has only been explored partially. [5] (Kumar et al) presents bidirectional LSTM with attention layers and Doc2vec embeddings in identifying the clickbaits. They focus on the multimedia dataset. However, attention played a significant role in identifying the clickbait . (Oliver et al, 2022) [4] performs spoiler generation and used scraped dataset from facebook and reddit and applied roberta-squad2 and T5 models to help users [3]. This paper describes the differences between extractive QA and abstractive QA and apply both approaches on different datasets and has concluded that better model would be dependent on the dataset selected. (Matthias et al, 2022) [3] used Question- Answering based approach for the spoiler generation and used Roberta and deberta for spoiler generation. For the classification of spoiler into phrase, passage and multi, the paper use three different approaches, one vs one, one vs rest and multi and concludes that in all the three approaches, transformers based classification works better as compared to feature based classification using models like SVM, naïve based and more. In our research, we have kept only multi class approach I.e we will classify the spoiler between three types- phrase, passage or multipart. For the spoiler generation, (Matthias et al, 2022) [3] used QA based approach on transformers like ROberta, Deberta but the results for passage were not as good as for phrase.Also, the paper scope was limited to only phrase and passage but in our research we also focussed on multi part.

3 Research Methodology

3.1 Spoiler Type Classification

The training dataset contained 1367 phrases, 1274 passages, and 559 multi-type spoilers. The dataset was highly imbalanced, with a small number of multi-type spoilers. This imbalance could lead to models making mistakes in determining the type of spoiler. To address this, we fine-tuned a robust model that could classify spoilers of all types.

- We began by extracting the relevant attributes, namely the PostText and targetParagraphs, from the training data. We then combined the postText with the targetParagraphs using a separator token and performed tokenization with a maximum length of 512 using the RoBERTa tokenizer. This process resulted in word-level tokens that were subsequently input into the fine-tuned RoBERTa base model.

- We trained the RoBERTa model with a learning rate of 5e-5, batch size of 16, and weight decay of 0.01 for 10 epochs. We stopped the training when the loss stopped decreasing, and achieved an F1 score of 70.3%.

- We also fine-tuned several other models, including Bert-base uncased, DistilBERT, GPT3, and Longformer. However, the best results were achieved with RoBERTa base, which yielded a 70.3% F1 score.

- For GPT3, we trained two models: Ada and Curie. Ada was trained using the OpenAI API with two different kinds of prompts from the training dataset, prompt1 and prompt2. Curie was trained with prompt3, which was similar to prompt1. The F1 score for GPT3 Curie was close at 70.02%, however GPT3 Ada was not able to achieve greater than or equal to 70% F1 score. It is worth mentioning that the GPT3 Curie model was trained for only 4 epochs due to its high computational requirements.

Overall, the results of our experiments show that RoBERTa base is the most effective model for spoiler detection.

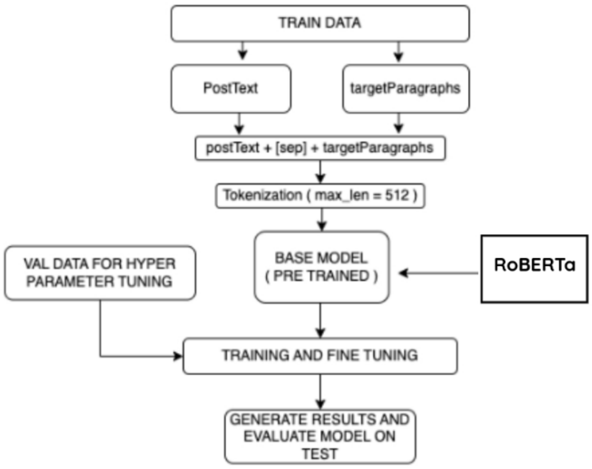

TRAIN DATA
PostText
targetParagraphs
postText + [sep] + targetParagraphs
Tokenization (max_len = 512)
VAL DATA FOR HYPER PARAMETER TUNING
BASE MODEL (PRE TRAINED)
RoBERTa
TRAINING AND FINE TUNING
GENERATE RESULTS AND EVALUATE MODEL ON TEST

Figure 1: Final Architecture for Spoiler Type Classification

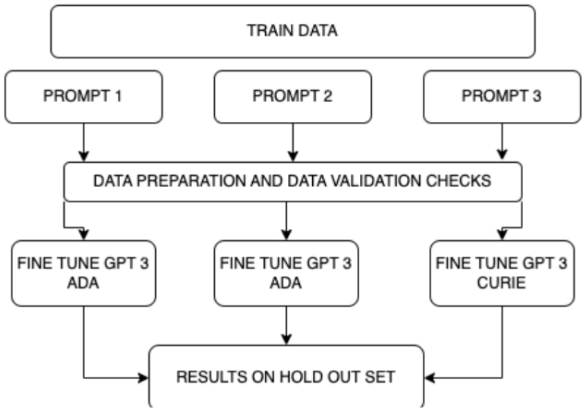


Figure 2: GPT3 Architecture for Spoiler Type Classification

3.2 Spoiler Generation

3.2.1 Text Pre-processing

The text pre-processing step involves preprocessing of spoilers and obtaining the spoilers tags and tokens and mapping them in between the target-Paragraphs for QA based apprach. For conditional based generation mapping of tags was not done with targetParagraphs since it was not extractive answer based approach.

3.2.2 Model Architecture

- In our initial approach, the models were applied to the entire dataset, constituting the baseline. To enhance our results, we partitioned the dataset into three distinct subsets: phrase, passage, and multi. Subsequently, we applied appropriate models to each subset. Our baseline architecture was further refined through the process of fine-tuning, incorporating BERT, Roberta, BART, T5 base, and T5 long models[1, 2]. Among these models, T5 long demonstrated exceptional efficacy in generating spoilers for passages and multipart texts, while T5 base exhibited optimal performance in generating spoilers for phrases. These findings contribute to the improvement and effectiveness of our research methodology.

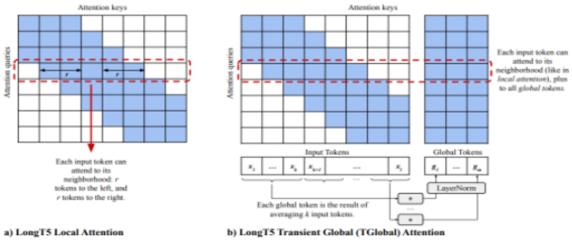
- (Matthias et al, 2022) [3] presents effective question answering based approach for extracting spoilers which extracts spoilers as answers for postext as questions and treating targetParagraphs as context. Hence for question answering based approach, we fine-tuned BERT, RoBERTa to view how effective it is to treat postTexts as questions and extract spoilers as answers keeping targetParagraphs as context. This question answering based approach was easily compared with Conditional generation approach by T5 base in which we obtained better results for T5 base. This suggested us to improve further on conditional generation.

- We observed the length of the targetParagraphs was maximum near to 4500 words in dataset and this length was hard to be fitted into models such as BERT, Roberta base, BART [6] and T5 base since they were operating on 512 max length or 1024 max length. T5 long can perform on 16384 token size which is greater than the maximum target paragraphs length.

- Hence, we applied long T5 model which provided us with longer input and attention span as compare to T5. The attention span refers to the range of input tokens that each token attends to during the attention computation. With more attention layers and a larger attention span, T5 Long can potentially capture longer-range dependencies in the input text. Additionally, In T5 Base, the model has 12 attention layers, while T5 Long has 24 attention layers. This difference in the number of attention layers affects the model's capacity to capture dependencies and relationships between words in the input text.

- T5 long focuses on the local attention and transient global attention which is useful in processing long texts. Local attention focuses on attention of a token from the nearby tokens e.g in a input of length l, nearby r tokens are selected and processed in a sliding window manner which gives the complexity as presented in equation 1 and diagram of local attention mechanism in presented in figure. For transient global attention -TGlobal, each token attention considers nearby as well as global attention. If l is the input sequence and k is the block size, TGlobal attention introduces a block of l l/k additional attention key-value pairs to calculate on top of Local Attention (l input tokens, attending to l/k global tokens; represented by the right most rectangle in Figure 2.b, hence for input sequence length l, complexity is represented by equation 2 .

Local attention complexity: $O(l \times r)$
TGlobal attention complexity : $O(l(r + \frac{l}{k}))$


Attention keys
Attention queries
r
r
Each input token can attend to its neighborhood: r tokens to the left, and r tokens to the right.
Attention keys
Attention queries
Each input token can attend to its neighborhood (like in local attention), plus to all global tokens.
Input Tokens
x_1 ... x_k x_{k+1} ... x_l
Global Tokens
g_1 ... g_m
LayerNorm
+
+
Each global token is the result of averaging k input tokens.

a) LongT5 Local Attention

b) LongT5 Transient Global (TGlobal) Attention

LongT5: Efficient Text-To-Text Transformer for Long Sequences (Mandy et. al)

Figure 3: Local attention and Transient global attention

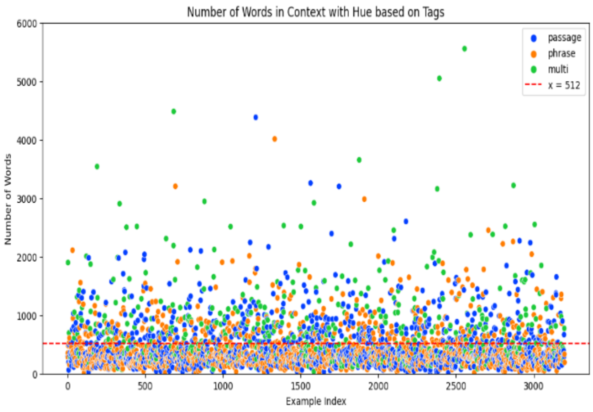


Figure 4: Number of words in the samples

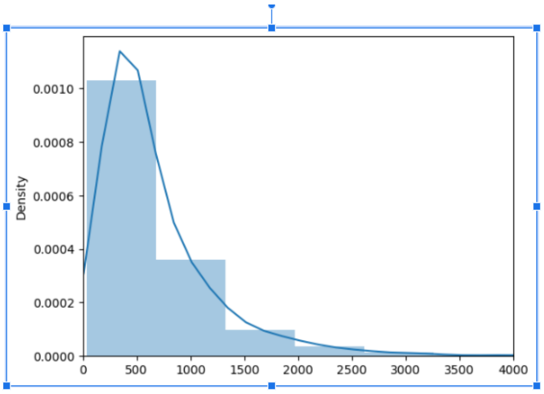


Figure 5: density vs number of words

3.3 Results

3.3.1 Spoiler Type Classification

We achieved the best F1 score with RoBERTa base model for classifying the spoilers among phrase, passage and multi. F1 score of all the models are presented in the Table 1. Second best results were achieved with GPT3 which were near to RoBerta and that was 70.02 %. This should be noted that GPT3 was trained on only 4 epochs due to computationally expensive operations. We compared improved model with the baseline Bi-LSTM model and found that F1 score was improved from 49% to 70.38% which indicates that RoBERTa is more suitable for classifying the spoilers. Results derieved were only on multi class classification approach not on one vs rest or one vs one. In research (Matthias et al, 2022) [3], author achieved the F1 score of 73% which was around 3% higher than our results.

Table 1: Models vs number of words

	Number of words
Most LLMS	512
T5	**512**
Long T5	16384
LED	16384

Table 2: Results of Spoiler Classification

	F1 Score
BERT-base*	61.16
ROBERTA-base*	**70.38**
DistilBERT*	65.01
GPT-3 (Ada - Prompt 1)*	68.50
GPT-3 (Ada - Prompt 2)*	67.22
GPT-3 (Curie - Prompt 3*)	70.02

3.3.2 Spoiler Generation

As we applied different models on segragated datasets of spoilers that are phrase, passage and multi dataset. We obtained different metrics for phrase, passage and multi. Our evaluation metrics for the datasets is presented in table 2,3 and 4.

In baseline, we were considering the whole dataset and trying to generate spoilers without segregation and hence baseline results were not good. We improved that in our miletsone3 models and results can be compared as shown in the table 5.

- For phrase we achieved best results with Long former encoder decoder model which were near to T5 Base. Bleu score of LED is 0.55. This should be noted that for phrases T5 base is better than T5 long since its attention window is smaller than T5 long. Hence for one-two extraction of words it performs better.

- For passages, we achieved best results with long T5 model which are near to 90%. This is because of transient global attention of long T5 model which can take more number of words in attention window and hence increasing the attention span of long T5. One of the reasons is also because long T5 can operate on higher number of words as compare to T5 base, BERT, RoBERTa base that we used.

- For multi, we achieved best results with long T5 model which were 0.81 bleu score. The reasons were also transient global attention and longer sequences processing in this case.

We compared our results with paper (Matthias et al, 2022) [3] published in ACL and our model outperforms the passage and multi part of spoiler generation but lags behind in case of phrase spoiler generation. Results can be shown in table 6.

Table 3: Results of PHRASE Spoiler Generation

	BLEU	Meteor	BERT Score
BERT*	0.21	0.48	60.34
ROBERTA*	0.31	0.59	71.51
T5 Base*	0.550	**0.61**	**97.05**
LED*	**0.553**	0.576	96.24
Long T5*	0.547	0.599	96.68

Table 4: Results of PASSAGE Spoiler Generation

	BLEU	Meteor	BERT Score
BERT	0.13	0.27	20.63
MiniLM	0.17	0.27	24.25
T5 Base*	0.77	0.83	96.67
LED*	0.853	0.885	97.70
Long T5*	**0.891**	**0.906**	**98.16**

Table 5: Results of MULTI Spoiler Generation

	BLEU	Meteor	BERT Score
BERT	0.032	0.12	14.05
MiniLM	0.025	0.12	13.86
T5 Base*	0.59	0.69	93.18
LED*	0.754	0.825	95.79
Long T5*	**0.818**	**0.862**	**96.57**

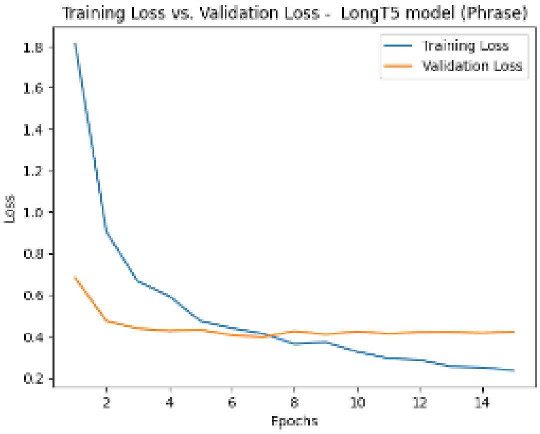


Figure 6: Training Loss vs Validation Loss (Long T5 - Phrase)

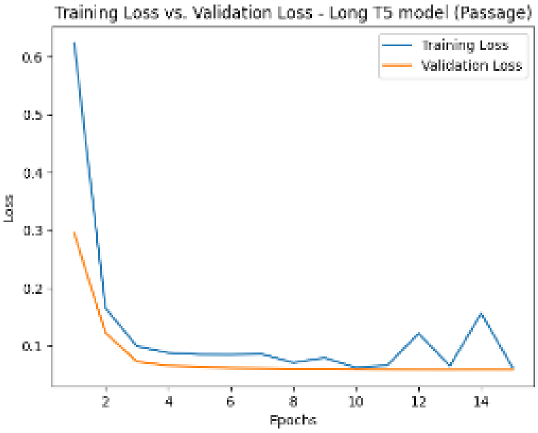


Figure 7: Training Loss vs Validation Loss (Long T5 - Passage)

4 Discussion and Error Analysis

4.1 Interpretation of Results

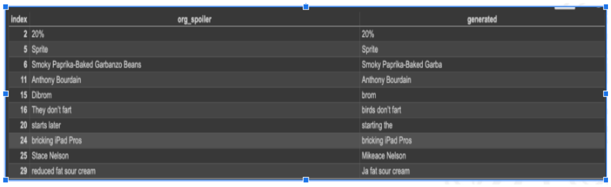
The results of the training and validation loss provide invaluable understanding of how our fine-tuned Long T5 models perform and behave when dealing with various types of spoilers, including phrase, passage, and multi-spoilers.

In the case of the 'phrase' spoiler type, the training loss consistently decreases over 15 epochs, demonstrating effective learning. However, the validation loss does not exhibit a comparable consistent decline. Instead, it initially decreases but later displays slight fluctuations, indicating the model's difficulty in dealing with the inherent complexity of generating 'phrase' spoilers.

When it comes to the 'passage' spoiler type, the model demonstrates a more consistent pattern of performance. The training and validation losses both experience substantial decreases, suggesting that the model has effectively learned and can apply this knowledge to unfamiliar data

In the case of the 'multi' spoiler type, there is a notable decrease observed in both the training and validation losses.

4.2 Variation in Performance



index	org_spoiler	generated
2	20%	20%
5	Sprite	Sprite
6	Smoky Paprika-Baked Garbanzo Beans	Smoky Paprika-Baked Garba
11	Anthony Bourdain	Anthony Bourdain
15	Dibrom	brom
16	They don't fart	birds don't fart
20	starts later	starting the
24	bricking iPad Pros	bricking iPad Pros
25	Stace Nelson	Mikeace Nelson
29	reduced fat sour cream	Ja fat sour cream

Figure 8: Long T5 Spoiler Generation Results - Phrase

Figure 8 illustrates the outcomes obtained from

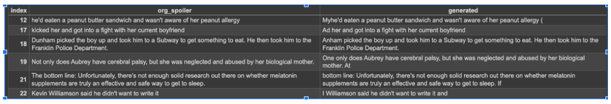
Table 6: ACL Paper (Matthias et al, 2022) vs Ours

	Phrases		**Passage**		**Multi**	
	ACL	**Ours**	**ACL**	**Ours**	**ACL**	**Ours**
BLEU	**68.80**	55.37	31.44	**89.18**	NA	**81.85**
Meteor	**67.94**	57.64	46.06	**90.69**	NA	**86.20**

the LongT5 model for phrase-type spoilers. Although the model displays an understanding of the fundamental content, consistent replication is not achieved. While certain phrases like "Anthony Bourdain" and "reduced fat sour cream" are accurately transcribed, the model falls short in other cases. For instance, phrases like "Smoky Paprika-Baked Garbanzo Beans" are shortened, and "They don't fart" is transformed into "birds don't fart."

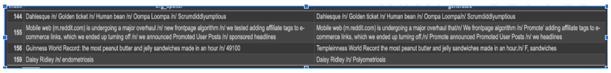
Regarding the performance of the T5 Long model on passage-type spoilers (Figure 9), it effectively preserves the core elements and narrative structure of the original text. Despite occasional minor distortions or syntax errors, the model generally maintains coherence.

In the case of multi-type spoilers (Figure 10), the T5 Long model performs well in capturing the essence of the original text and generating coherent spoilers. Essential phrases such as "Dahlesque," "Golden ticket," "Human bean," "Oompa Loompa," "Scrumdiddlyumptious," and "Daisy Ridley" are successfully replicated.



index	org_spoiler	generated
12	he'd eaten a peanut butter sandwich and wasn't aware of her peanut allergy	Myhe'd eaten a peanut butter sandwich and wasn't aware of her peanut allergy (
17	kicked her and got into a fight with her current boyfriend	Ad her and got into a fight with her current boyfriend
18	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.	Anham picked the boy up and took him to a Subway to get something to eat. He then took him to the Franklin Police Department.
19	Not only does Aubrey have cerebral palsy, but she was neglected and abused by her biological mother.	One only does Aubrey have cerebral palsy, but she was neglected and abused by her biological mother. At
21	The bottom line: Unfortunately, there's not enough solid research out there on whether melatonin supplements are truly an effective and safe way to get to sleep.	bottom line: Unfortunately, there's not enough solid research out there on whether melatonin supplements are truly an effective and safe way to get to sleep. If
22	Kevin Williamson said he didn't want to write it	I Williamson said he didn't want to write it and

Figure 9: Long T5 Spoiler Generation Results - Passage



index	org_spoiler	generated
144	Dahlesque /n/ Golden ticket /n/ Human bean /n/ Oompa Loompa /n/ Scrumdiddlyumptious	Dahlesque /n/ Golden ticket /n/ Human bean /n/ Oompa Loompa/n/ Scrumdiddlyumptious
155	Mobile web (m.reddit.com) is undergoing a major overhaul /n/ new frontpage algorithm /n/ we tested adding affiliate tags to e-commerce links, which we ended up turning off /n/ we announced Promoted User Posts /n/ sponsored headlines	Mobile web (m.reddit.com) is undergoing a major overhaul that/n/ We frontpage algorithm /n/ Promote' adding affiliate tags to e-commerce links, which we ended up turning off./n/ Promote announced Promoted User Posts /n/ we headlines
156	Guinness World Record: the most peanut butter and jelly sandwiches made in an hour /n/ 49100	Templeinness World Record the most peanut butter and jelly sandwiches made in an hour./n/ F, sandwiches
159	Daisy Ridley /n/ endometriosis	Daisy Ridley /n/ Polyometriosis

Figure 10: Long T5 Spoiler Generation Results - Multi

5 Conclusion

When it comes to the phrase type of spoiler, adopting a question-answering (QA) approach may yield better results. We achieved the best bleu score for LED model which is 0.55. On the other hand, for passage and multi spoilers, employing a summarized extraction by long T5 model method proves to be highly effective. We achieved best bleu score by T5 which is 0.89 for passage and 0.81 for multi part types of spoilers. However, to achieve state-of-the-art (SOTA) performance on these spoiler types, it is essential to utilize a Long T5 model that has been trained on a large-scale question answering dataset.

By incorporating longer generation models and transient global attention mechanisms of long T5, we have successfully achieved SOTA results in handling passage and multi spoiler types.

6 Contribution

Contribution for milestone1, 2 and 3 are given in the table 7. Some of them just includes the research work that was performed on different models, models implemented and presentation contributions.

References

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

[2] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences, 2022.

[3] Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. Clickbait spoiling via question answering and passage retrieval, 2022.

[4] Oliver Johnson, Beicheng Lou, Janet Zhong, and Andrey Kurenkov. Saved you a click: Automatically answering clickbait titles, 2022.

[5] Vaibhav Kumar, Dhruv Khattar, Siddhartha Gairola, Yash Kumar Lal, and Vasudeva Varma. Identifying clickbait. In *The 41st International ACM SIGIR Conference on Research &amp Development in Information Retrieval*. ACM, jun 2018.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

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

Table 7: Contribution

	Task
Milestone1	Read the research papers mentioned in the related work of this report.
Milestone1	Data Analysis
Milestone1	Discussions with team about the approaches need to consider
Milestone1	Added slides in ppt – data analysis, problem statement, objectives, feature extraction
Milestone 2	Report for complete spoiler generation part
Milestone 2	Bert for spoiler generation
Milestone 2	Text preprocessing
Milestone 2	Tried to ensemble BERT + T5
Milestone 3	Text preprocessing for QA based Approach
Milestone 3	Bert fine tuned on phrase generation
Milestone 3	Roberta fine tuned on phrase
Milestone 3	Tried roberta for passage and multi but it was computationally expensive
Milestone 3	MiniLM [8] for passage
Milestone 3	Researched on Pegasus[9] and Luke model for passage and multi
Milestone 3	PPT

[8] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[9] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.