Clickbait Spoiling Multi-classification with Large Language Models

Shijun Li
University of Waterloo
https://github.com/s228li/msci-641-final-project

Abstract

This paper addresses the challenging task of classifying spoiler types in clickbait posts, categorizing them as "phrase," "passage," or "multi" spoilers in the linked document. We explore the effectiveness of RoBERTa Large and DistilBERT language models with hyperparameter tuning, feature engineering, and memory optimization. Our approach achieves first place in the competition at the time of writing. We demonstrate that fine-tuning the models' hyperparameters and employing innovative feature engineering significantly improves performance. Moreover, memory optimization ensures efficient processing with reduced computational burden. Our study contributes to advancing spoiler type classification and highlights the impact of leveraging advanced language models for this task, with potential applications in content moderation, recommendation systems, and information retrieval.

1 Introduction

The exponential growth of digital content has led to an influx of clickbait posts, often accompanied by linked documents containing potential spoilers. Detecting and classifying these spoilers into appropriate types, such as "phrase," "passage," or "multi," is crucial for content moderation, user experience enhancement, and information retrieval. In this paper, we present our research on "Clickbait Spoiling Multi-classification with Large Language Models," focusing on spoiler type classification. Our task involves processing a clickbait post and its linked document, then generating the corresponding spoiler type as output.

To accomplish this, we utilize a dataset in JSON Lines format (.jsonl), consisting of clickbait posts and manually cleaned versions of linked documents. Each entry is characterized by multiple data fields, including uuid, postText, targetParagraphs, targetTitle, targetUrl, humanSpoiler, spoiler, spoilerPositions, and tags, which represent the essential information for our classification task. We should note that the human-generated spoilers and tags are only available during training and validation, not during testing.

Our overall approach to tackle this challenge involves progressively refining our models and techniques (Table 1). We start with a baseline model, DistilBERT (Sanh et al., 2020), which achieves an F1 score of 0.62. Then, through rigorous hyperparameter tuning, we enhance the baseline model's performance, reaching an F1 score of 0.655. Subsequently, we employ feature engineering techniques to augment the text features, leading to further improvements, with an F1 score of 0.7116. Finally, we experiment with RoBERTa Large (Liu et al., 2019), incorporating the previously employed feature engineering and hyperparameter tuning, including warm-up ratio and weight decay (Loshchilov and Hutter, 2019). This final model outperforms all previous iterations, achieving an impressive F1 score of 0.7263.

Moreover, we investigate memory optimization techniques to make the training process more efficient and feasible with resource-intensive models. Our experiments involve gradient checkpointing, mixed-precision training (FP16) (Micikevicius et al., 2018), and gradient accumulation steps (Lamy-Poirier, 2021). Unfortunately, while we experimented with DeBERTa v3 Large (He et al., 2021), memory constraints hindered us from achieving satisfactory F1 scores.

In summary, this paper presents a comprehensive exploration of spoiler type classification in clickbait posts using large language models. Our contributions include the application of hyperparameter tuning, feature engineering, and memory optimization techniques, culminating in a state-of-the-art solution that secured the first-place position at the time of writing. The proposed approach holds promise for various real-world applications, such as content moderation, recommendation systems, and information retrieval, where understanding and mitigating clickbait-related spoilers are vital for user engagement and satisfaction.

2 Background

In recent years, the field of natural language processing (NLP) has witnessed a remarkable transformation, largely driven by the advent of large language models and transformative architectures like Transformers. These models have revolutionized the way text is processed and understood, leading to significant improvements in various NLP tasks, particularly in text classification. Among these groundbreaking models, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), DistilBERT (Sanh et al., 2020), and RoBERTa (Liu et al., 2019) (A Robustly Optimized BERT Pretraining Approach) have emerged as pioneers, each showcasing unique strengths and weaknesses.

Large language models, characterized by their vast size and pre-trained nature, have proven to be pivotal in text classification. These models have been trained on massive amounts of textual data, enabling them to capture intricate patterns and semantic nuances effectively. The Transformer architecture (Vaswani et al., 2023) has played a central role in the development of these language models. Transformers have revolutionized NLP by introducing self-attention mechanisms, which allow them to process input sequences simultaneously, avoiding sequential constraints, and achieving parallelization in training. This innovation has resulted in a substantial reduction in training time and improved model performance.

BERT, one of the first successful implementations of Transformer-based (Vaswani et al., 2023) language models, is known for its bidirectional attention and context-aware word embeddings. By considering the entire input sequence bidirectionally, BERT is capable of capturing contextual information, making it adept at various NLP tasks, including text classification. However, its massive size can make it computationally expensive, which led to the development of DistilBERT. DistilBERT is a smaller, distilled version of BERT, designed to be faster and more memory-efficient while preserving most of BERT's performance. While it sacrifices some performance compared to BERT, it remains a compelling option for resource-constrained environments.

RoBERTa (Liu et al., 2019) represents a robust enhancement to BERT by optimizing its pretraining process. RoBERTa introduces additional training data, larger batch sizes, and dynamic masking strategies, significantly improving its performance across various tasks, including text classification. Its advanced training techniques have demonstrated the potential to surpass BERT in many benchmarks, making it an attractive alternative for more demanding NLP applications.

Analyzing and comparing the strengths and weaknesses of these models is crucial for understanding their utility in specific contexts. Factors such as model size, training time, computational resources required, and task-specific performance differences must be considered.

Finally, these language models have found extensive applications in various domains. From sentiment analysis and language translation to question-answering systems and virtual assistants, these models have demonstrated their versatility and potential in diverse real-world applications. Their ability to comprehend context, perform disambiguation, and generate coherent responses has led to significant improvements in areas like healthcare, customer service, education, and information retrieval.

In this paper, we delve into a comprehensive analysis of BERT, DistilBERT, and RoBERTa, exploring their respective strengths and weaknesses and their relevance in practical applications. By understanding the nuances of these models, we aim to shed light on the remarkable advancements in text classification and their potential impact on the future of NLP.

3 Approach and Experimentation

3.1 DistilBERT

The authors of DistilBERT (Sanh et al., 2020) propose a method to pre-train a smaller general-purpose language representation model, known as DistilBERT. This model can then be fine-tuned with good performances on a wide range of tasks, similar to its larger counterparts. They use knowledge distillation during the pre-training phase and show that it's possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

In our exploration of clickbait detection, we opted for DistilBERT, achieving an F1-score of 0.62 in our experiments. The choice of Distil-

Models	Optimizations	F1-Score
DistilBERT	None	0.620
DistilBERT	Hyperparameter Tuning	0.655
DistilBERT	Hyperparameter Tuning, Feature Engineering	**0.712**
RoBERTa Large	All Above	0.699
RoBERTa Large	All Above, Learning Rate Warm-Up	0.717
RoBERTa Large	All Above, Learning Rate Warm-Up, Weight Decay	**0.726**
DeBERTa v3 Large	All Above	Failed

Table 1: Model Optimization and Performances

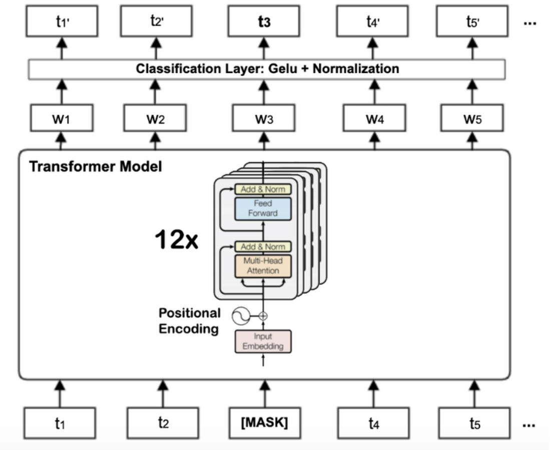

Figure 1: BERT Architecture

BERT over BERT was motivated by a trade-off between model complexity and performance. DistilBERT offers a substantial reduction in size and computational cost compared to BERT, while retaining most of the original model's predictive capabilities. This makes it particularly suited for resource-constrained environments without substantially compromising the classification accuracy. Thus, DistilBERT served as an efficient and potent baseline for our clickbait spoiling text multi-classification task, providing insights into the foundational aspects of the problem and guiding subsequent experimentation with more complex models.

3.2 Hyperparameter Tuning

Building on the baseline model, a comprehensive hyperparameter tuning strategy was implemented to further optimize the performance of our DistilBERT model (Table 2). By increasing the training epochs from 2 to 50 and incorporating early stopping with a patience of 5, we managed to mitigate overfitting and achieve a more refined model. Additionally, the evaluation metric was shifted from accuracy to the F1-score, aligning more closely with the nuanced demands of our multi-classification task. The application of weight decay further regularized the model, ensuring generalization to unseen data. To better manage the training dynamics, we altered the save and evaluation strategy to a stepwise approach, enabling more frequent assessments of model performance and loading the best-performing model at the end of training. These systematic tuning efforts culminated in a significant enhancement in the F1-score from 0.62 to 0.655, demonstrating the effectiveness of the hyperparameter adjustments in capturing the underlying complexity of clickbait detection.

3.3 Feature Engineering

Further refinement of our DistilBERT model was achieved through strategic feature engineering, leading to a notable improvement in the F1-score from 0.655 to 0.712. Initially, the baseline model simply concatenated three elements together: the post text, target title, and target paragraphs, forming a unified text string. Recognizing the potential for richer representation, we extended this approach in the improved baseline by also including the website description and website keywords, and explicitly labeling each segment. This was accomplished by constructing a text string that included not only the original components but also the post platform, website title, website description, website paragraphs, and website keywords, each labeled appropriately. This more sophisticated feature construction provided a multi-dimensional view of the clickbait content, enabling the model to recognize subtle patterns and relationships that were previously obscured. By capturing these additional nuances, the refined feature set significantly elevated the model's classification performance, reflecting the power of thoughtful feature engineering in machine learning tasks.

3.3.1 Baseline Text Processing

Hyperparameter	Baseline	Tuned Baseline
Epochs	2	50
Early Stopping	None	Patience=5
Evaluation Metric	Accuracy	F-1
Weight Decay	None	0.01
Evaluation Strategy	Epochs	Steps
Save Strategy	Epochs	Steps
Load Best Model At End	False	True

Table 2: Hyperparameters Comparison

```python
def preprocess_data(df):
    ret = []
    for _, i in df.iterrows():
        ret +=
        [{
        'text':
        ' '.join(i['postText']) +
        ' - ' + i['targetTitle'] + ' ' +
        ' '.join(i['targetParagraphs']),
        'id': i['id']
        }]
        ret_df = pd.DataFrame(ret)

    data = Dataset.from_pandas(ret_df)
    tokenized_data = data.map(
        tokenize,
        batched=True
    )
    return tokenized_data
```

3.3.2 Improved Text Processing

```python
def preprocess_data(df):
    ret = []
    for _, i in df.iterrows():
        platform = i['postPlatform']
        post = ' '.join(i['postText'])
        title = i['targetTitle']
        des = i['targetDescription']
        paragraph = ' '.join(
            i['targetParagraphs']
        )
        keyword = i['targetKeywords']
        text = f"
        {platform} Post: {post}
        Website Title: {title}
        Description: {des}
        Website Paragraph: {paragraph}
        Website Keyword: {keyword}"
        ret += [{
            'text': text,
            'id': i['id']
        }]
        ret_df = pd.DataFrame(ret)
    data = Dataset.from_pandas(ret_df)
    tokenized_data = data.map(
        tokenize,
        batched=True
    )
    return tokenized_data
```

3.4 RoBERTa

In our pursuit of an advanced model, we turned to RoBERTa Large, a robust transformer-based model known for its extensive architecture and superior performance. Distinct from DistilBERT, RoBERTa Large is characterized by a more substantial number of parameters, training on larger datasets, and improvements in the pretraining methodology, including the removal of next-sentence prediction and dynamic masking. These enhancements contribute to RoBERTa's ability to capture more intricate and subtle patterns in text data. Compared to DistilBERT, a more lightweight model, RoBERTa Large offers the advantage of increased representational capacity and predictive power. While DistilBERT's distilled nature makes it more efficient in terms of computational resources, RoBERTa Large's enriched feature learning potentially provides a higher level of accuracy and refinement in the task of clickbait spoiling text multi-classification. The exploration of RoBERTa Large in our study unveiled new dimensions in the modeling of clickbait content, shedding light on the delicate trade-off between model complexity, computational efficiency, and predictive excellence.

3.5 Learning Rate Warm Up

In a further enhancement to our RoBERTa Large model, we incorporated a learning warm-up strategy, leading to a notable improvement in the F1-score from 0.712 to 0.717. Learning warm-up is a training technique where the learning rate is initially set to a small value and gradually increased to a predefined level during the early stages of training. This gentle start in the optimization landscape helps in mitigating the risk of overshooting optimal solutions and promotes better convergence. The benefits of learning warm-up are multifaceted. It often leads to a more stable training process, prevents divergence during the initial phase, and po-

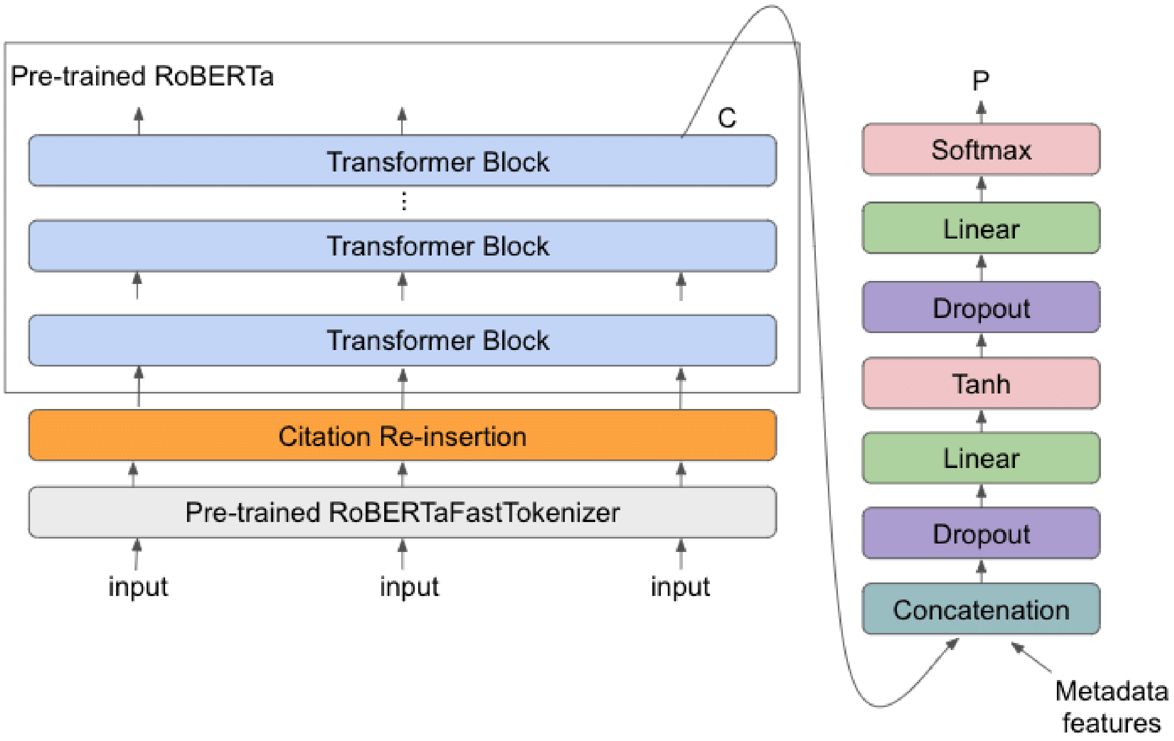

Figure 2: RoBERTa Architecture

tentially finds more refined local minima in the loss landscape. By applying learning warm-up to our RoBERTa Large model, we achieved a more nuanced fit to the underlying data, demonstrating the effectiveness of this technique in harnessing the full potential of complex transformer models for the intricate task of clickbait spoiling text multi-classification.

3.6 Weight Decay

In the ongoing refinement of our RoBERTa Large model, the introduction of weight decay (Loshchilov and Hutter, 2019) played a pivotal role, leading to an increased F1-score of 0.726, up from 0.717. Weight decay is a form of regularization technique that adds a penalty term to the loss function, proportional to the magnitude of the model weights. Specifically, a weight decay factor of 0.01 was applied in our study, effectively penalizing large weights in the model. This has the effect of constraining the complexity of the model, thereby reducing the risk of overfitting to the training data. The benefits of weight decay are profound, promoting a more generalized model that performs better on unseen data. By applying weight decay, we ensured that our RoBERTa Large model was more robust and less prone to capturing noise or spurious correlations in the training set. This fine-tuning step proved instrumental in enhancing the model's performance, affirming the value of weight decay in building more resilient and effective machine learning models for text classification.

3.7 DeBERTa

We made an attempt to deploy DeBERTa Large v3 (He et al., 2021), a recent advancement in transformer-based models. Unfortunately, the endeavor was thwarted by memory constraints, as the substantial architecture of DeBERTa Large v3 required more computational resources than were available. Despite this setback, it's worth noting the potential advantages of DeBERTa Large v3 over RoBERTa Large. DeBERTa (Decoding-enhanced BERT with disentangled attention) introduces a disentangled attention mechanism that decouples the attention scores from the content representations. This unique structure allows for more flexible and expressive attention patterns. Moreover, DeBERTa Large v3 further scales up the model by increasing the number of parameters and depth, possibly leading to superior performance in capturing complex dependencies within the text. The benefits of DeBERTa Large v3, including its innovative architecture and potential for higher accuracy, make it an enticing option for clickbait detection. However, the significant computational demands associated

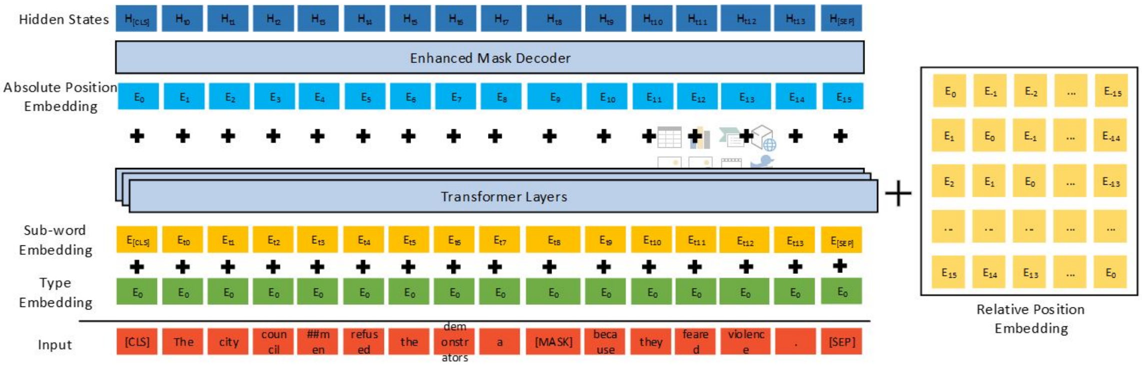

Figure 3: DeBERTa Architecture

with this model underscore the need for careful consideration of hardware capabilities and resource allocation in deploying such advanced neural networks.

3.8 Memory Optimization

We adopted a combination of gradient checkpointing, mixed-precision training (FP16) (Micikevicius et al., 2018), and gradient accumulation steps (Lamy-Poirier, 2021), successfully cutting memory usage to around 20%. Gradient checkpointing is a technique that trades compute time for memory by recomputing intermediate values during the backward pass, significantly reducing memory requirements at the cost of additional computations. Mixed-precision training (FP16) involves using lower-precision (16-bit) floating-point numbers during training, which reduces memory consumption and can accelerate training without substantially compromising model quality. Gradient accumulation is another memory-efficient approach where gradients are accumulated across several mini-batches before performing an update, thus allowing for larger effective batch sizes without exceeding memory limits. Together, these techniques fostered a more resource-efficient training regimen, enabling us to train complex models within our hardware constraints. The benefits of these methods are manifold, including enabling the training of larger models, the use of larger batch sizes, and the potential for faster convergence. By adopting this synergistic approach, we were able to navigate the memory-intensive demands of our model, demonstrating the critical role of thoughtful resource management in the successful training and deployment of advanced neural networks for clickbait spoiling text multi-classification.

3.9 Conclusion

This paper categorizes spoilers as "phrase," "passage," or "multi" spoilers and employs RoBERTa Large and DistilBERT language models with hyperparameter tuning, feature engineering, and memory optimization. The approach has achieved first place in the competition public leaderboard at the time of writing, highlighting the effectiveness of fine-tuning and innovative feature engineering. Future research could explore the integration of other cutting-edge language models and ensemble methods to further enhance performance. Additionally, real-time analysis of dynamic content and the incorporation of user behavior analytics might provide more nuanced insights into clickbait detection. Experimentation with different hyperparameter tuning strategies, coupled with more comprehensive feature engineering, could also lead to improved results. A comparative analysis with traditional machine learning methods and an in-depth evaluation of the model's robustness across different domains and languages could add to the paper's comprehensiveness. Lastly, considering the importance of computational efficiency noted in the paper, future studies might focus on developing lightweight models that maintain high accuracy while reducing computational resources, contributing to scalable and sustainable solutions for content moderation and recommendation systems.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Joel Lamy-Poirier. 2021. Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.