Clickbait Spoiling : Multiclass classification (Project Report)

Lamia Islam
Master Data Science
University of Potsdam
lamia.islam@uni-potsdam.de

Abstract

Clickbait headlines are designed to captivate readers by strategically omitting key details, often leading to ambiguous or misleading interpretations. This project addresses spoiler-type classification—predicting whether a clickbait post requires clarification via a *phrase*, *passage*, or *multi*-part spoiler—using the Webis Clickbait Spoiling Corpus 2022. Our approach combines domain-adapted (DA) word embeddings to capture clickbait semantics with three models of increasing complexity: a baseline Logistic Regression, a Bi-directional Long Short-Term Memory Network (Bi-LSTM) with attention mechanism, and a transformer-based ModernBERT model. Our work explores how advanced natural language processing (ANLP) can be leveraged to unravel the deceptive tactics of clickbait, thereby promoting clearer communication and more informed online discourse.

1 Introduction

Clickbait headlines have become ubiquitous in online media, designed to entice users to click through to articles by deliberately withholding key information or creating a "curiosity gap" (Loewenstein, 1994). These headlines often use sensationalist language, hyperbolic phrases, or deliberately ambiguous language to capture attention. While effective at generating traffic, clickbait can lead to user frustration when the content fails to deliver on the headline's implicit promise, contributing to information pollution and undermining trust in digital media (Molyneux and Coddington, 2020).

Clickbait spoiling addresses this issue by automatically providing the missing information that a clickbait headline intentionally omits. Rather than forcing users to click through to potentially disappointing content, spoiling systems reveal the key information, the "spoiler", that was deliberately withheld (Hagen et al., 2022). However, not all clickbait can be spoiled in the same way. Some require just a short phrase, while others need a more detailed passage to adequately explain the withheld information. Some complex cases may even require multiple spoilers to fully address the promise of clickbait (Fröbe et al., 2023).

In this project, we focus on the task of clickbait spoiler type classification—determining whether a given clickbait post requires a phrase, passage, or multi-part spoiler. This classification serves as a crucial first step in a complete clickbait spoiling system, allowing subsequent processes to generate appropriate spoilers based on the identified type. By correctly classifying the spoiler type needed, we can more effectively counter clickbait tactics and provide users with the information they seek without unnecessary clicks.

This task has practical implications for improving information access and digital literacy. When implemented in browser extensions or social media platforms, automated spoiling systems could help users quickly assess whether content is worth their attention, reducing the effectiveness of misleading clickbait strategies. Additionally, the ability to identify and classify clickbait contributes to broader efforts to improve online content quality and transparency, building upon earlier work in clickbait detection (Potthast et al., 2016; Chakraborty et al., 2016).

In this project, we explored three models of increasing complexity to tackle this classification challenge: a baseline Logistic Regression model using TF-IDF features, a Bidirectional Long Short-Term Memory (BiLSTM) network with attention mechanism to better capture contextual information, and a transformer-based ModernBERT model. We also investigated the impact of domain-adapted word embeddings specifically tuned to clickbait linguistic patterns.

Through this work, we aim to develop an effective approach to clickbait spoiler type classification while gaining deeper insights into the linguistic characteristics that distinguish different forms of clickbait, ultimately contributing to more transparent and honest online communication.

2 Related Work

The foundation of our approach builds upon recent advancements in clickbait spoiling and text classification techniques.

(Hagen et al., 2022) introduced clickbait spoiling as a two-step process: first classifying the spoiler type (phrase, passage, or multipart), then extracting the appropriate spoiler. Their evaluation demonstrated that state-of-the-art question answering models, particularly DeBERTa-large, outperformed passage retrieval methods for generating both phrase and passage spoilers. Their findings highlight the importance of spoiler type classification as a first step, as it can significantly improve the effectiveness of the spoiling process when done accurately.

(Fröbe et al., 2023) formalized this task through the SemEval 2023 Clickbait Spoiling competition, where RoBERTa-based classifiers achieved the highest accuracy (approximately 74%) for spoiler type classification. Their work established standard evaluation metrics that we adopt in our methodology, creating a framework for consistent comparison with existing approaches. The strong participation of the competition (30 teams) demonstrated the growing interest in this research direction.

For classification specifically, (Sharma et al., 2023), presented an information condensation approach that achieved the highest accuracy in the spoiler type classification. Their two-step method first used contrastive learning to identify the most relevant paragraphs in an article, then applied DeBERTa for classification based on this condensed information. This approach demonstrated the value of focusing on relevant content rather than processing entire articles, aligning with our goal to develop efficient classification methods.

Our technical approach draws from three additional works representing increasing levels of model complexity. (Sarma et al., 2018) presented domain-adapted word embeddings that combine generic and domain-specific representations using Canonical Correlation Analysis (CCA). This technique is particularly relevant for capturing the unique linguistic patterns of clickbait content that may not be well represented in generic embeddings.

(Liu et al., 2024) proposed an optimized BiLSTM network with attention for news text classification, capturing bidirectional dependencies and focusing on the most relevant textual features. Their architecture serves as our intermediate model, balancing complexity and performance.

Finally, (Warner et al., 2024) introduced ModernBERT, an advanced bidirectional encoder that improves upon previous transformer-based models in terms of both performance and efficiency. The effectiveness of the model provides a strong foundation for our most advanced classification approach.

Our work extends these foundations by: (1) investigating whether domain-adapted embeddings can better capture clickbait-specific linguistic patterns, (2) implementing a BiLSTM with attention mechanism as an intermediate solution between simple logistic regression and the complex transformer models, and (3) evaluating ModernBERT's capabilities specifically for spoiler type classification. Through this combination, we aim to advance classification performance while investigating the relationship between model complexity and effectiveness for this task.

3 Task Formalization

The clickbait spoiler type classification task can be formally defined as a supervised multiclass classification problem. Given a clickbait post and its linked article, the objective is to classify the post into one of three categories: phrase, passage, or multi-part spoiler. This classification is an essential first step in the broader clickbait spoiling process, determining the appropriate structure of information to extract in a subsequent spoiler generation phase (Hagen et al., 2022).

3.1 Problem Definition

Formally, we define the task as follows:

Let $\mathcal{D} = \{(p_i, a_i, y_i)\}_{i=1}^{N}$ be a dataset consisting of N samples, where for each sample i:

- p_i represents the clickbait post text
- a_i represents the linked article text
- $y_i \in \mathcal{Y} = \{\text{phrase}, \text{passage}, \text{multi}\}$ is the ground truth label indicating the type of spoiler required

The goal is to learn a classification function $f : P \times A \to Y$ that maps a clickbait post and its linked article to the appropriate spoiler type.

3.2 Model Implementation

Our approach employs three primary modeling strategies of increasing complexity:

1. **Logistic Regression with TF-IDF (Baseline)**: A traditional approach using term frequency features, providing interpretability and serving as a performance benchmark for more sophisticated models.

2. **BiLSTM with Attention**: A sequential neural network that processes text bidirectionally while focusing on relevant portions through an attention mechanism to capture contextual dependencies in clickbait headlines.

3. **ModernBERT**: A fine-tuned transformer-based model leveraging pre-trained contextual embeddings to capture semantic nuances between clickbait headlines and their corresponding spoilers.

3.3 Evaluation Framework

Following the PAN Clickbait Challenge at SemEval 2023 (Fröbe et al., 2023), the primary evaluation metric for this task is balanced accuracy across all three classes, which accounts for potential class imbalance. Secondary metrics include precision, recall, and F1 score for each individual spoiler type, providing a more detailed understanding of model performance.

4 Data

This section outlines the dataset utilized for our clickbait spoiler classification task, along with relevant statistics and analysis of its characteristics.

4.1 Dataset Description

For our work, we used the Webis Clickbait Spoiling Corpus 2022 (Hagen et al., 2022). The corpus provides a standard split with 3,200 posts for training (80%) and 800 posts for validation (20%). While the complete corpus contains an additional 1,000 test posts, these were not available for our project work and were reserved for future evaluation purposes. Our model development and experimentation were therefore conducted using only the training and validation sets.

4.2 Spoiler Types

A significant feature of this dataset is the categorization of spoilers into three types, illustrated in Figure 1 :

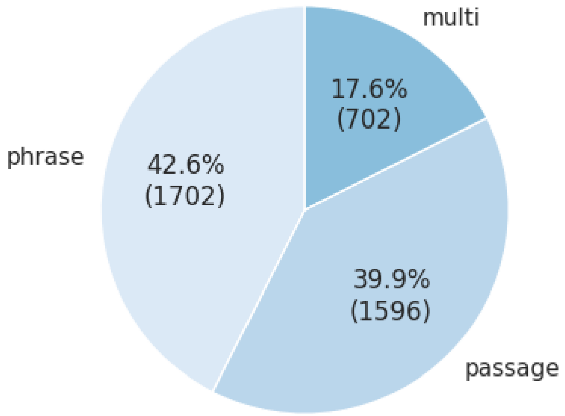

Figure 1: Distribution of different spoiler types

- Passage spoilers: Longer spans consisting of one or a few sentences, with an average length of 24.1 words.

- Phrase spoilers: Short spans consisting of a single word or phrase from the linked document, with an average length of 2.8 words.

- Multipart Spoilers: Consisting of more than one non-consecutive phrase or passage from the linked document, with an average length of 33.9 words.

4.3 Data Statement

Following Bender and Friedman (2018) data statement framework, here are the additional analysis of the corpus:

CURATION RATIONALE The corpus was created to support research on automatically spoiling clickbait posts by generating short texts that satisfy the curiosity induced by the post. The clickbait posts were primarily sourced from various social media platforms: Twitter (47.5%), Reddit (36%), and Facebook (16.5%) and supplemented with posts from the Webis-Clickbait-17 corpus that were manually spoiled by the creators (Hagen et al., 2022).

LANGUAGE VARIETY The corpus consists exclusively of English language content.

SPEAKER DEMOGRAPHICS The demographic information of the original authors of the clickbait posts is not available, as they were collected from public social media platforms and news websites.

ANNOTATOR DEMOGRAPHICS The annotation was primarily performed by one main annotator with verification by two additional experts among the co-authors of the original paper (Hagen et al., 2022). No specific demographic information about the annotators is provided.

SPEECH SITUATION The clickbait posts represent public social media communication aimed at attracting clicks.

TEXT CHARACTERISTICS The corpus includes clickbait posts that typically employ linguistic techniques designed to create curiosity gaps, such as sensationalism, teasers, and cataphors. The linked documents are primarily news articles, product reviews, and blog posts from various domains.

RECORDING QUALITY N/A.

OTHER N/A.

PROVENANCE APPENDIX N/A.

4.4 Sample Data Points

To better illustrate the nature of the dataset, Table 1 presents examples of clickbait posts with their associated spoilers across different spoiler types, drawn from the training set.

5 Experiments

While Section 3 presented the formal approach and high-level modeling strategies, this section elaborates on the specific architectural details, implementation choices, and experimental setup to ensure reproducibility.

5.1 Logistic Regression

Our baseline logistic regression model employs multiple feature extraction approaches:

- Feature Variants: We implemented and tested three TF-IDF feature variants:
  - Standard: Features extracted from the combined title and paragraph text with a maximum of 10,000 features
  - Separate: Title and paragraph texts processed independently, then combined via horizontal stacking
  - N-gram: Utilizing unigrams, bigrams, and trigrams (n-gram range 1-3)
- Word Embedding: Beyond TF-IDF, we also implemented weighted document embeddings where title and article text vectors are combined with weights of 0.7 and 0.3 respectively.
- Hyperparameter Configuration: The model uses balanced class weights, L2 regularization with C=0.1, and the LBFGS solver with 1000 maximum iterations.

This array of feature engineering approaches allows us to identify the most effective text representation for the clickbait-spoiling task while maintaining the interpretability advantages of the logistic regression framework.

5.2 BiLSTM with Attention

The BiLSTM model architecture includes several technical optimizations:

- Word Embedding: We initially observed severe overfitting. Incorporating pre-trained Google News word embeddings reduced this issue.
- LSTM Structure: A single-layer bidirectional LSTM with configurable hidden dimensions (default: 124 units) and attention mechanism to focus on the most informative words.
- Training Process: The model employs AdamW optimizer (lr=3e-4, weight_decay=8e-4), gradient clipping at 0.5, and a learning rate scheduler that reduces the rate by 50% after 3 epochs without improvement.

5.2.1 ModernBERT

Given the limited performance of previous approaches, we experimented with ModernBERT models across various hyperparameter configurations.

- Model Variants: We experiment with both ModernBERT-base and ModernBERT-large variants from the answerdotai repository.
- Tokenization Strategy: Inputs are structured with the title repeated three times, followed by a [SEP] token and article content, then tokenized and padded to 128 tokens.

Clickbait Post	Spoiler	Type
"NASA sets date for full recovery of ozone hole"	"2070"	Phrase
"Just how safe are NYC's water fountains?"	"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."	Passage
"A Harvard nutritionist and brain expert says she avoids these 5 foods that weaken memory and focus."	"1. Added sugar", "2. Fried foods", "3. High-glycemic-load carbohy- drates", "4. Alcohol", "5. Nitrates"	Multipart

Table 1: Examples from the Webis Clickbait Spoiling Corpus 2022

- Training Configuration: Fine-tuning uses a batch size of 16, learning rate of 8e-5 with linear warmup (ratio 0.1), and weight decay of 0.01.

- Optimization Process: The model uses early stopping with a patience of 3 evaluations, evaluating performance every 100 steps, and saving checkpoints with a maximum of 2 saved models.

- Resource Management: For environments with limited GPU memory, we implement optional mixed precision training (fp16) and gradient accumulation.

5.3 Evaluation Protocol

Building on the evaluation framework outlined in Section 3.3, our experimental procedure includes:

- **Metrics Implementation:** We use scikit-learn's implementations of balanced_accuracy_score and classification_report for consistent evaluation across models.

- **Result Saving:** All evaluation results are saved as JSON files with timestamps for traceability.

- **Model Versioning:** Best-performing models are saved with their corresponding tokenizers, vocabulary mappings, and configuration parameters to enable direct reuse.

This experimental setup enables systematic comparison of the three modeling approaches while providing sufficient detail for future replication of our results.

6 Results

This section presents the experimental results of our clickbait spoiling models across three approaches: TF-IDF with logistic regression, BiLSTM with attention, and ModernBERT. We analyze performance through balanced accuracy, precision, recall, and F1 scores, with attention to the three spoiler classes.

6.1 Logistic Regression

Our baseline logistic regression with TF-IDF features showed moderate performance with signs of overfitting. The basic TF-IDF with 10,000 features achieved a training balanced accuracy of 0.6901 but only 0.5209 on validation data. Reducing the feature space to 1,500 yielded similar results (0.7078 training, 0.5223 validation). The most effective configuration was the separate TF-IDF approach, which treated titles and paragraphs as distinct feature sets, achieving a validation balanced accuracy of 0.6018. This suggests that preserving the distinct linguistic patterns between clickbait headlines and spoiler text is important for effective classification.

6.2 BiLSTM with Attention

In our BiLSTM experiments, we initially observed severe overfitting. Incorporating pre-trained Google News word embeddings reduced this issue but still didn't yield competitive performance. The final configuration (vocabulary size: 20,000, sequence length: 100, embedding dimension: 300, hidden dimension: 128, dropout: 0.65) achieved a balanced accuracy of 0.5438 and a weighted F1 score of 0.5499. The class-specific F1 scores were 0.6037 for phrase spoilers, 0.5240 for passage spoilers, and 0.4821 for multi-element spoilers.

Despite its theoretical capacity to capture long-range text dependencies, the BiLSTM model underperformed compared to the baseline logistic regression model, suggesting it struggles to model the complex relationships needed for distinguishing between spoiler types, particularly for multi-element spoilers.

6.3 ModernBERT Models

Given the limited performance of previous approaches, we experimented with transformer-based ModernBERT models across various hyperparameter configurations.

To enable systematic analysis, we curated a structured results dataset from our evaluations, selecting representative configurations that allowed for meaningful parameter comparisons. All visualizations and analyses presented derive from this organized collection.

6.3.1 Overall Performance Distribution

Figure 2 shows the distribution of performance metrics across different spoiler classes for our ModernBERT experiments. The distributions show that ModernBERT models significantly outperformed our previous approaches, with balanced accuracy values centered around a median of 0.65 and extending to a maximum of 0.71. This represents substantial improvement over the best TF-IDF model (0.60) and BiLSTM model (0.54). Similarly, the F1 score distribution demonstrates consistent performance advantages, with a median of 0.67 and best results reaching 0.73, compared to 0.61 and 0.55 for TF-IDF and BiLSTM respectively. The distribution of metrics indicates that ModernBERT models provide robust performance even across varying hyperparameter settings.

6.3.2 Learning Rate Analysis

Our learning rate optimization followed a two-phase approach: first testing a logarithmic range (5e-04, 5e-05, 5e-06) to identify promising magnitudes, then conducting fine-grained experiments from 1e-05 to 9e-05.

Table 2 presents ModernBERT-base performance across learning rates while keeping other parameters constant (sequence length=128, batch size=16, epochs=3).

Performance improved steadily from 1e-05 to 8e-05, then declined at 9e-05, suggesting optimal gradient updates need to be sufficiently large to escape suboptimal regions but not so large as to cause training instability.

Learning Rate	Balanced Accuracy
1e-05	0.5469
3e-05	0.6542
5e-05	0.6800
8e-05	0.6940
9e-05	0.6646

Table 2: Learning Rates Comparison

6.3.3 Sequence Length Optimization

Figure 3 demonstrates that sequence length is a critical parameter. A maximum length of 128 tokens provided optimal balance between sufficient context and avoiding noise. Longer sequences (256 tokens) typically led to performance degradation, while shorter sequences (64 tokens) failed to capture enough contextual information, suggesting clickbait spoiling requires moderate context but suffers from information dilution with excessively long sequences.

6.3.4 Training Duration Effects

As shown in Figure 4, three epochs provided optimal performance for most configurations. Beyond this point, validation loss increased while training loss continued to decrease, indicating overfitting. This relatively short optimal training duration suggests the models quickly learn relevant patterns, and additional training primarily reinforces dataset-specific biases rather than improving generalization.

6.3.5 Impact of Model Size

As Figure 5 illustrates, ModernBERT-large consistently outperformed its base counterpart. With the optimal learning rate of 8e-05, the large model achieved a balanced accuracy of 0.7077 versus 0.6940 for the base model. The performance advantage was most pronounced for multi-element spoilers, with the large model achieving an F1 score of 0.6642 compared to 0.6520 for the base model, suggesting that additional parameters help capture the nuanced linguistic patterns required for complex spoiler classification.

6.4 Class-specific Performance

Figure 6 shows consistent patterns in class-specific performance. Phrase spoilers (class 0) were easiest to predict, with the highest F1 scores across all models. Passage spoilers (class 1) showed moderate performance, while multi-element spoilers

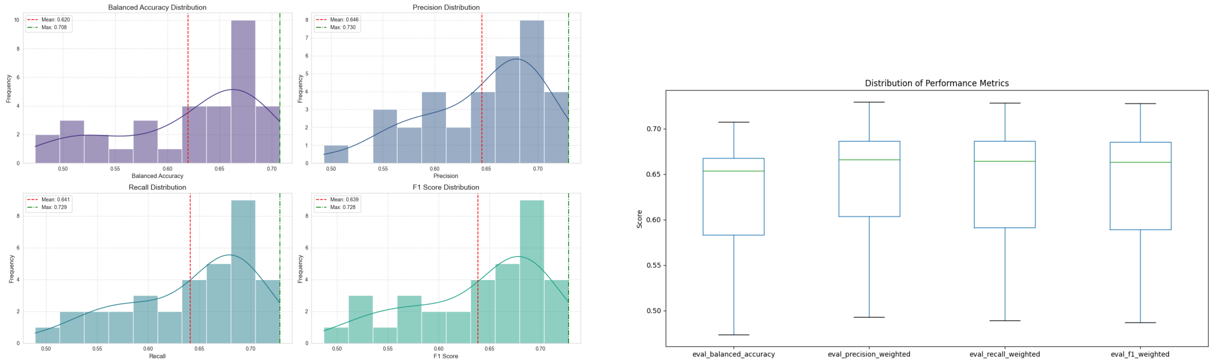

Figure 2: Performance metrics for different ModernBERT configurations

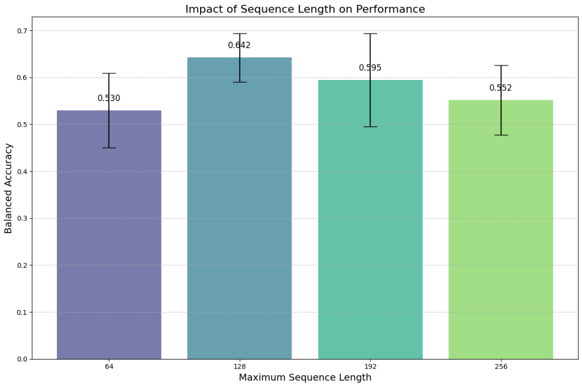

Figure 3: Impact of maximum sequence length on model performance

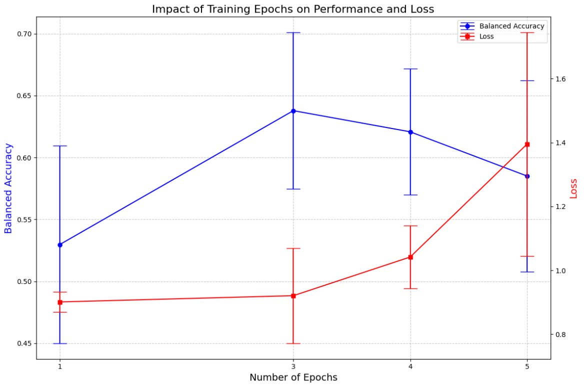

Figure 4: Performance across different training durations

(class 2) proved most challenging. ModernBERT-large substantially narrowed these performance gaps compared to simpler approaches.

6.5 Error Analysis

DRAFT Our error analysis revealed several linguistic patterns associated with misclassifications:

Complex multi-element spoilers: The models struggled most with spoilers requiring understanding of multiple components and their relationships. For example, spoilers containing both numerical data and named entities were often misclassified as passage spoilers.

Ambiguous phrase boundaries: When phrase spoilers contained multiple clauses or dependent phrases, models sometimes misclassified them as passage spoilers, suggesting difficulty in identifying precise spoiler boundaries.

Implicit information: Spoilers relying on inference or implicit knowledge were frequently misclassified. For instance, headlines asking "Who won?" were challenging when the spoiler contained contextual information beyond just the winner's name.

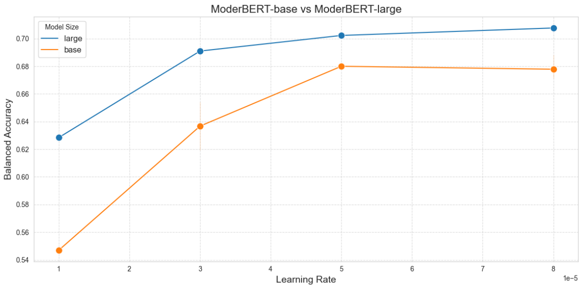

Figure 5: Performance comparison between ModernBERT-large and ModernBERT-base

Domain-specific terminology: Technical or domain-specific terminology in spoilers (e.g., sports statistics, medical terms) led to higher error rates, indicating the models' sensitivity to vocabulary distribution.

Semantic overlap between classes: Cases where spoilers contained characteristics of multiple classes (e.g., a brief passage with numerical elements) showed higher error rates, reflecting the inherent ambiguity in spoiler categorization.

These linguistic challenges were progressively better handled by more sophisticated models, with

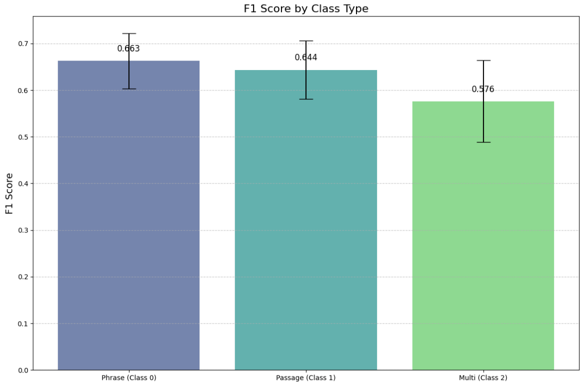

Figure 6: Class-specific F1-scores for ModernBERT models

ModernBERT-large showing the most robust performance across these difficult cases.

6.6 Model Comparison Summary

The results in Table 3 demonstrate that transformer-based models significantly outperform traditional approaches for clickbait spoiling, with ModernBERT-large providing the most robust performance across all metrics and spoiler types. The substantial improvement in class 2 performance is particularly noteworthy, as this represents the most challenging category of spoilers.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Maik Fröbe, Tim Gollub, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. Semeval-2023 task 5: Clickbait spoiling. In *17th International Workshop on Semantic Evaluation (SemEval 2023)*, pages 2278–2289. Association for Computational Linguistics.

Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. 2022. Webis clickbait spoiling corpus 2022.

Bingyao Liu, Junming Huang, Jiajing Chen, Yuanshuai Luo, Rui Wang, and Jianjun Wei. 2024. Optimizing news text classification with bi-lstm and attention mechanism for efficient data processing. *arXiv preprint arXiv:2409.15576*.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Logan Molyneux and Mark Coddington. 2020. Aggregation, clickbait and their effect on perceptions of journalistic credibility and quality. *Journalism Practice*, 14(4):429–446.

Prathusha K Sarma, Yingyu Liang, and William A Sethares. 2018. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 51–59, Melbourne, Australia. Association for Computational Linguistics.

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 (SemEval-2023)*, pages 1878–1889, Toronto, Canada. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Model	**Balanced Accuracy**	**F1-Score Phrase**	**F1-Score Passage**	**F1-Score Multi**
Logistic Regression	0.6018	0.6321	0.6175	0.5458
BiLSTM	0.5438	0.6037	0.5240	0.4821
ModernBERT-base	0.6940	0.7141	0.7294	0.6520
ModernBERT-large (best)	0.7077	0.7413	0.7418	0.6642

Table 3: Results Comparison for different Models