Clickbait Spoiling Subtask 1 : Multiclass classification

Akash Palh
University of Potsdam
palh@uni-potsdam.de

Abstract

In today's digital landscape, clickbait's captivating allure is undeniable, driving our investigation into Subtask 1 of the SemEval 2023 PAN Clickbait Challenge: classifying clickbait articles. Leveraging Long Short-Term Memory (LSTM) networks, enhanced with embedding and bidirectional layers, we aim to dissect clickbait content with precision, aligning with audience preferences to sustain narrative engagement.

Our methodology is anchored in rigorous data preprocessing, employing regular expressions, text cleaning, and TF-IDF vectorization to refine textual data. This groundwork paves the way for our LSTM model's optimization, where hyperparameter tuning—adjusting learning rate, batch size, and architecture—plays a pivotal role in bolstering classification accuracy. Validation metrics such as accuracy, precision, recall, and the F1 score, derived from a confusion matrix, offer a comprehensive assessment of our model's capabilities.

By analyzing the Webis Clickbait Spoiling Corpus 2022, we venture beyond mere classification, exploring ethical dimensions of clickbait. This phenomenon, often criticized for misleading audiences, prompts us to advocate for integrity in online information dissemination. Our endeavor not only confronts practical challenges associated with clickbait but also fosters an ethical dialogue, encouraging content creators to prioritize transparency and honesty.

This study, therefore, not only advances the academic conversation on natural language processing and clickbait detection but also highlights the significance of ethical considerations in digital content creation. By enhancing clickbait classification, we contribute to a more informed and ethically responsible digital environment, ensuring content integrity while catering to audience curiosity.

1 Introduction

Clickbait articles draw in readers with catchy headlines that spark curiosity instead of giving detailed summaries. The aim here is to classify a title and article into one of three spoiler types: "phrase," "passage," or "multipart," as shown in Figure 1. Identifying the type of spoiler helps improve how spoilers are made by tailoring content to what audiences want and like, leading to better engagement through accurate tagging. This approach keeps the story's suspense by sharing just enough information to keep interest alive, allowing for content that fits the context and audience needs. Moreover, it helps plan better content strategies to create excitement and collects feedback for improving future work, ensuring a story experience that meets audience expectations while keeping the story's essence intact.

In the rapidly evolving digital landscape, the intriguing phenomenon of clickbait presents a unique challenge and opportunity for natural language processing (NLP) and machine learning technologies. Our project, set within the framework of SemEval 2023 Task 5, targets the multifaceted task of classifying clickbait articles by their spoiler type. Leveraging the innovative capabilities of Long Short-Term Memory (LSTM) networks, enhanced with embedding layers and bidirectional processing, our model endeavors to dissect the intricate patterns underlying clickbait content, aiming for an accurate categorization that resonates with the dynamic preferences of online audiences.

The motivation behind this study is rooted in the dual nature of clickbait: while it effectively garners viewer attention, its potential to mislead raises significant ethical concerns. Our approach seeks to illuminate the mechanisms of clickbait, offering insights that could guide the creation of more transparent and informative online content. Preliminary

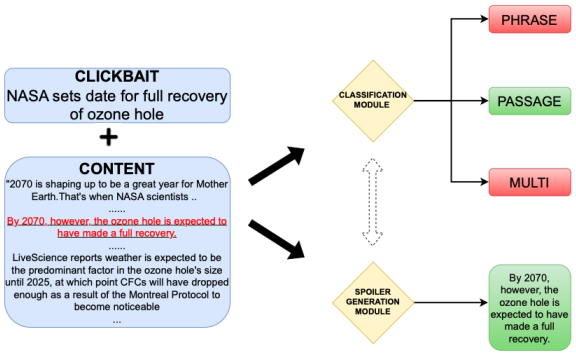

Figure 1: The text highlighted in red represents the spoiler, categorized as a "passage." The dotted arrow indicates a dependency between the two tasks.

results of our model reveal its capacity to classify clickbait inputs into their respective categories accurately. However, the performance, as measured by conventional accuracy metrics, has been modest. This limitation is primarily attributed to the constraints of our dataset—characterized by its limited size—and the stipulation against employing pre-trained models for this classification task.

As we navigate these challenges, our study not only contributes to the academic discourse surrounding NLP and clickbait detection but also engages with the broader ethical dialogue concerning digital content creation. By advancing our understanding of clickbait classification, we aspire to foster a more ethically responsible and engaging online environment, where content integrity and audience interest are harmoniously aligned.

2 Related Work

In our study, we adopt an innovative approach that significantly differs from the strategies utilized by (Wangsadirdja et al., 2023) and (Sharma et al., 2023). Whereas their research primarily emphasizes question-answering models and information condensation methods for mitigating clickbait, our methodology centers on the application of an advanced bidirectional Long Short-Term Memory (LSTM) model. This distinction is crucial; the bidirectional LSTM facilitates a deeper comprehension of text sequences by analyzing contextual relationships from both directions, an essential feature for the precise identification and classification of clickbait content.

Our novel contribution is particularly evident in the sophisticated pre-processing techniques we apply to text data. This includes stripping URLs, numbers, and stopwords from the content, thus streamlining the data fed into our neural network. Moreover, we innovatively aggregate the processed titles and paragraphs into a unified textual entity, enhancing the bidirectional LSTM's ability to effectively ascertain and categorize the relevant aspects of clickbait spoiling.

By diverging from the reliance on rephrasing titles for question-answering frameworks or compressing information to underline key points, our strategy marks a pivotal shift in addressing clickbait spoiling. It leverages the full spectrum of capabilities offered by bidirectional LSTMs, enabling an all-encompassing analysis of text data. This approach not only broadens the discourse initiated by (Wangsadirdja et al., 2023) and (Sharma et al., 2023) but also illustrates an alternative, possibly more effective, route to comprehending and addressing the nuances of clickbait narratives.

3 Task Formalisation

In addressing the issue of categorizing clickbait into the classifications of phrase, passage, or multipart, we delineate this endeavor as a task under the supervised learning paradigm. The essence of this task revolves around the discernment of clickbait content by its nature of spoilage, with the overarching aim to cultivate a predictive model adept at distinguishing among these specified categories accurately.

This process initiates with a comprehensive data preprocessing phase, incorporating text cleansing and its transformation into a numerical format suitable for machine learning models. Such preprocessing is critical to render the data amenable to the algorithmic processing that follows.

We employ a Bidirectional Long Short-Term Memory (Bi-LSTM) network for this task, an architecture particularly efficient in analyzing and learning from textual data in both forward and backward directions. The application of a Bi-LSTM model facilitates the interpretation of preprocessed textual content into a structured form, which is then learned by the model.

The efficacy of the Bi-LSTM model in accurately categorizing clickbait content is gauged through several evaluation metrics, including accuracy, the F1 score, precision, and recall. These metrics furnish a quantifiable insight into the model's proficiency in addressing the defined task.

Subsequent to the evaluation, the visualization of outcomes via graphical plots provides a deeper understanding and interpretation of the model's performance. This visual representation not only elucidates the model's effectiveness in clickbait classification but also assists in pinpointing areas for improvement.

4 Data

In the dataset designated for our project, the division between the training and validation datasets is predefined, with the training set comprising **3,200** entries and the validation set containing **800** entries. Each entry is uniquely identifiable by a universally unique identifier (uuid) and encompasses a variety of attributes including `postId`, `postText`, `postPlatform`, `targetParagraphs`, `targetTitle`, `targetDescription`, `targetKeywords`, `targetMedia`, `targetUrl`, `provenance`, `spoiler`, `spoilerPositions`, and `tags`, presenting a comprehensive overview of each post's characteristics.

Descriptive Statistics Overview:

`Post Texts`: The dataset features **3,186** distinct post texts, indicating minimal repetition among entries. `Platforms`: The source platforms are primarily three, with Twitter emerging as the predominant platform. `Target` Paragraphs: Showcasing diversity, there are **3,184** unique paragraphs within the dataset. `Target Title`: Exhibits a unique title for almost every entry, underlining the dataset's diversity. `Target Description`: Out of **2,933** entries with descriptions, **2,813** are unique, though some entries are devoid of any description. `Target Keywords`: Among the entries containing keywords, **1,795** distinct sets are noted. `Target Media`: The dataset includes **2,663** unique media links, suggesting a variety of media content. `Target URL`: Demonstrates a high uniqueness rate among URLs provided in the dataset.

Sample Entries Illustration:

A post titled "5 Reasons Why You Should Not Miss The 'Friends' Reunion" from Twitter is tagged as a passage, with its target title being "Friends Reunion Special: What We Know So Far". Another Twitter post discusses the "Ozone Layer Hole Over Antarctica" with a corresponding target title "Ozone Hole Over Antarctica Larger Due to Stratospheric Weather Variations" and is tagged as a phrase. A third post from Twitter, "The Secret to Employee Happiness Isn't More Money", matches the sentiment of its target title "Money Isn't the Secret to Employee Happiness", also tagged as a phrase.

In aligning with the ethical and scientific validity guidelines proposed by Bender and Friedman, it's imperative to scrutinize the dataset across several dimensions such as the demographics represented, the data collection context (including the rationale behind curation), language variety, and annotator demographics. This analysis is crucial for identifying inherent biases, thereby ensuring the dataset's responsible usage. For instance, a dataset predominantly representing a single demographic may not yield equitable performance across diverse groups. Such scrutiny allows for a more nuanced understanding of the dataset's applicability and limitations, guiding toward its ethical employment in natural language processing tasks.

Given the necessity for rigorous data validation and bias scrutiny, the creation of models that circumvent these critical steps is discouraged. Building models on unscrutinized data risks embedding biases within the model's outcomes, potentially leading to inaccurate results when presented with well-curated data. Therefore, ensuring the integrity of the data through thorough validation and bias examination is paramount before its application in model training, ensuring both ethical and scientifically sound practices in natural language processing endeavors.

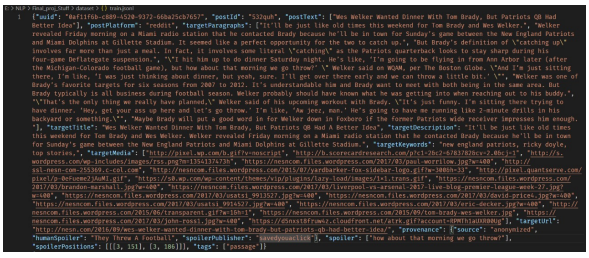

Figure 2: Dataset example

5 Experiments

This section outlines our experiment setup, informed by notable strategies from SemEval-2023 Task 5 on Clickbait Spoiling, aiming for reproducibility and further research enhancement.

5.1 Models and Methodologies

Jack-Ryder's Approach: Applies DeBERTa and Flan-T5 for transforming clickbait titles into questions and identifying relevant spoilers, emphasiz-

ing semantic similarity and QA functionalities (Wangsadirdja et al., 2023).

Give_GPT_2023's IC-based Framework: Utilizes RoBERTa and DeBERTa within an Information Condensation (IC) framework to streamline articles for accurate spoiler classification and generation, focusing on enhanced recall (Sharma et al., 2023).

5.2 Baselines

The above methods are compared against standard transformer models like BERT and RoBERTa, applied directly without task-specific optimizations, showcasing their innovative contributions.(Maik Fröbe, 2023)

5.3 Implementation Insights

The implementations, utilizing the Transformers library, emphasize minimal fine-tuning to harness the models' zero-shot capabilities for detecting and generating spoilers.

5.4 Evaluation Procedure

Jack-Ryder's Approach: Focuses on zero-shot learning efficiency, particularly in title rephrasing and spoiler extraction accuracy (Wangsadirdja et al., 2023).

Give_GPT_2023's IC-based Framework: Analyzes the IC strategy's effectiveness in spoiler classification and generation, using BLEU scores for accuracy measurement (Sharma et al., 2023).

Our approach employs LSTM networks, augmented with embedding layers for data handling and representation learning. Bidirectional LSTMs capture contextual nuances, with softmax activation enabling multi-class classification.

5.5 Preprocessing

We employ tokenization, regex for cleaning, and TF-IDF for numerical conversion. Standardization techniques include stop word removal, lowercase conversion, and contraction elimination.

5.6 Hyperparameter and Evaluation

Optimal hyperparameters such as learning rate, batch size, and LSTM architecture specifics are finely tuned. Model performance is evaluated using accuracy, precision, recall, and F1 score.

5.7 Optimization and Loss

Categorical cross-entropy is chosen for its multi-class suitability, alongside an optimizer for weight adjustment based on gradients.

5.8 Training and Final Evaluation

The model is trained to minimize loss, with continuous validation set monitoring. Post-training, it undergoes testing on a separate dataset to assess generalization capabilities, evaluated with predefined metrics for comprehensive performance insight.

This concise approach, from preprocessing to final testing, ensures our LSTM model is adept at accurately handling clickbait spoiling, delivering precise and contextually relevant predictions.

6 Results

This section delves into the experimental results, featuring visual representations and error analysis, with a focus on linguistic attributes.

6.1 Visualization and Summary Statistics

A word cloud visually ranks words from most to least frequent, offering insight into the dataset's linguistic composition.(Mohiuddin, 2024)

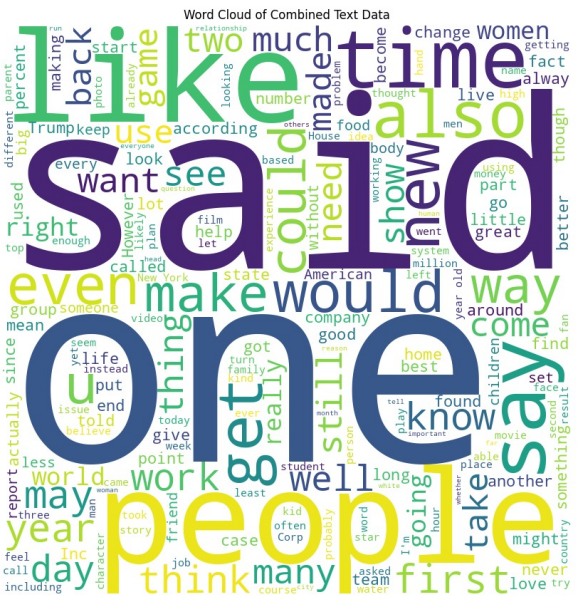

Figure 3: Word Cloud Visualization: Showcases the frequency distribution of words within the text data.

6.2 Descriptive Analysis

A detailed summary of text attributes, including length and variance, is tabulated, highlighting the dataset's textual diversity.(Mohiuddin, 2024)

Summary Statistics Table:

Statistic	Value
count	3200.000000
mean	300.269375
std	393.479901
min	4.000000
25%	130.000000
50%	207.000000
75%	367.000000
max	13824.000000

Table 1: Text Attribute Summary: Provides statistical measures like mean, median, and standard deviation for text lengths.

6.3 Text Length Distribution

A histogram displays the range and distribution of text lengths, illustrating variability across the dataset.

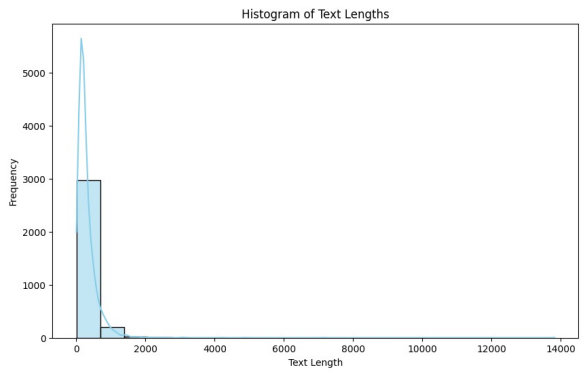

Figure 4: Text Length Histogram: Depicts the spread of text lengths within the dataset.

6.4 Frequency Analysis

Tables list the most common words, emphasizing the dataset's prevalent vocabulary.

Top 10 Most Frequent Words Table:

Word	Frequency
.	4688
,	3996
one	3702
like	3509
people	3171
would	2862
also	2833
—	2831
said	2552
get	2420

Figure 5: Frequent Words Analysis: Identifies the top N most frequent words, underlining lexical preferences.

6.5 Model Efficacy

Performance metrics are tabled, offering a quantitative assessment of the model's predictive accuracy.

Model Performance Table:

Metric	Value
Accuracy	0.437500
F1 Score	0.202899
Precision	0.812500
Recall	0.333333

Table 2: Performance Metrics Summary: Outlines the model's effectiveness through various evaluation metrics.

6.6 Classification Accuracy

A confusion matrix graphically represents the model's classification accuracy, detailing performance across categories.(Mohiuddin, 2024)

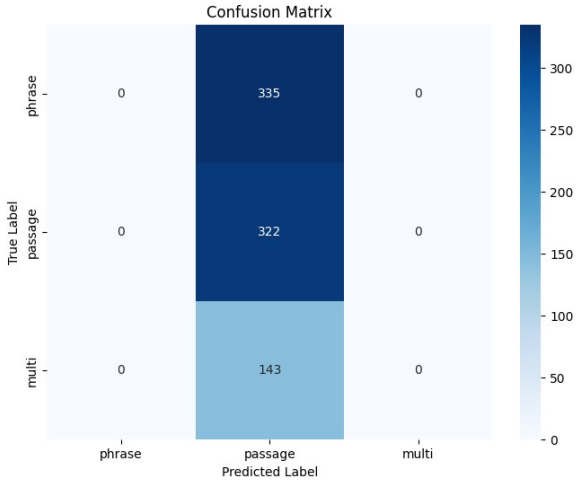

Figure 6: Confusion Matrix Overview: Visualizes true vs. predicted classifications, highlighting model precision.

6.6.1 Linguistic Dataset Characteristics

The dataset's linguistic features are examined, including language, structure, and preprocessing impact:

- **Language and Vocabulary:** The dataset, primarily in English, undergoes stop-words and URL removal, focusing on content-rich words.
- **Text Structure:** Combining titles and paragraphs creates diverse syntactic compositions, introducing varied sentence complexities.
- **Normalization Effects:** Text cleaning standardizes the dataset but may omit contextually significant details like specific stopwords or numbers.
- **Semantic Diversity:** The dataset's broad thematic range necessitates attention to topic diversity and potential biases.

6.6.2 Model Implementation Insights

The LSTM classifier's architecture and its ramifications are analyzed, identifying potential enhancement avenues:

- **Embedding and LSTM Layers:** The embedding layer enhances semantic learning, while the LSTM layer, crucial for sequential data, could benefit from a bidirectional approach for improved context capture.
- **Dropout and Connection:** Dropout helps prevent overfitting, and the fully connected layer bridges LSTM output to classification outputs, with RMSprop and CrossEntropyLoss optimizing for accuracy.

Improvement Strategies:

1. **Bidirectional LSTM Exploration** Could bolster contextual understanding.
2. **Hyperparameter Optimization:** Adjusting learning rates and layer specifics may enhance performance.
3. **Advanced Regularization:** Exploring dropout adjustments or other methods might further reduce overfitting.
4. **Evaluation Methods Expansion:** Incorporating additional metrics could provide deeper performance insights.
5. **Data Enrichment:** Data augmentation techniques might improve training robustness and model reliability.

7 Conclusion

In this endeavor, my primary contributions will revolve around the technical execution and thoughtful critique of our model's efficacy. With a strong foundation in programming, my focus will be directed towards the development and fine-tuning of the LSTM model's architecture and its evaluation strategies. Moreover, I aim to be an integral part of the project across all stages, dedicating myself to enhancing model accuracy through iterative improvements and keen analysis. My interpersonal skills will also be leveraged to foster a cohesive and driven team environment.

The completion of this project is anticipated to enrich our collective comprehension of LSTM networks, hyperparameter optimization, and the critical assessment of performance indicators. This experience has already broadened our perspective with global methodologies. Utilizing the specified tools and libraries, as discussed in my individual report, will further cement our practical skills, translating theoretical knowledge into actionable insights within a tangible framework.

In wrapping up, I intend to contribute significantly to data pre-processing and hyperparameter

adjustments, ensuring our model's robustness and precision. Being a proactive and supportive team member remains a top priority, as collaborative synergy is pivotal for the project's triumph.

8 Limitations and Ethical considerations

This research provides insightful contributions to the utilization of LSTM models in text classification tasks but faces several limitations and ethical challenges. Linguistically, the pre-processing approach—particularly the exclusion of stop-words and numerical data—might inadvertently dismiss the critical role these elements play in conveying the text's full semantic and syntactic depth. This could lead to a diminished comprehension of the content by the model. Additionally, the dataset's specificity poses concerns regarding its comprehensiveness and the representativeness of the samples used. Important details such as the dataset's licensing, collection methodology, and sharing permissions require clearer articulation to ensure ethical standards for data handling and dissemination in the scientific domain are met.

The potential societal repercussions and the dual-use nature of this research merit cautious reflection. The capabilities developed through this study may be repurposed in manners that could challenge privacy norms or magnify existing biases within the training data, thereby reinforcing societal disparities. This underscores the necessity for integrating fairness and anti-bias measures in NLP modeling to avert prejudiced outcomes. Challenges related to the project's scope, such as data breadth and the choice of evaluation metrics, reflect broader issues in fully grasping human language nuances and rendering model judgments transparent and answerable to end-users.(Bender and Friedman, 2018)

Another area of concern is the environmental toll of deep learning model training, which demands significant energy resources, contributing to the AI field's carbon footprint. Future endeavors should pursue more sustainable computational methods and weigh the balance between the environmental cost and the efficiency of model outcomes.

Emphasizing reproducibility, this study advocates for the open exchange of code and data resources to verify findings and encourage collective progress through shared knowledge. It also draws attention to the ethical duties of researchers to deliberate the wider effects of their investigations, including misuse risks and the commitment to advancing technology that is socially beneficial and minimizes adverse impacts. As NLP technology advances, it is crucial that ethical guidelines and conscientious research practices remain at the forefront of AI development and application.

References

Albatrot. 2023. The ethics of clickbait. Accessed: 2024-04-01.

Tushar Abhishek Radhika Mamidi Vasudeva Varma Anubhav Sharma, Sagar Joshi. 2023. Billy-batson at semeval-2023 task 5: An information condensation based system for clickbait spoiling. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

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.

Tim Gollub Matthias Hagen Martin Potthast Maik Fröbe, Benno Stein. 2023. Semeval-2023 task 5: Clickbait spoiling. *Association for Computational Linguistics*, Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Version 2.

Artur Jurk Martin Potthast Matthias Hagen, Maik Fröbe. 2022. Clickbait spoiling via question answering and passage retrieval. *Computation and Language (cs.CL)*.

Shahrukh Mohiuddin. 2024. Clickbait spoiling (task 1). *University of Potsdam - Department of 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.

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 (SemEval-2023)*, pages 1090–1095, Toronto, Canada. Association for Computational Linguistics.