Clickbait Spoiling Subtask 1 : Multiclass classification

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

Abstract

In this study, we tackle Subtask 1 of the SemEval 2023 PAN Clickbait Challenge, which involves the classification and generation of spoilers in clickbait articles. Our strategy employs Long Short-Term Memory (LSTM) networks with added embedding layers and bidirectional functionality to customize spoiler content according to the preferences and expectations of the audience, aiming for a compelling narrative experience. Our approach includes thorough data preprocessing, the tuning of hyperparameters, and the application of various validation metrics for assessing model performance. Leveraging the Webis Clickbait Spoiling Corpus 2022, we explore a range of analytical and data manipulation tools, underscoring our efforts to enhance the capabilities of LSTM models for effectively generating and classifying spoilers within the realm of clickbait content detection.

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 (spoiler generation - Subtask 2) 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.

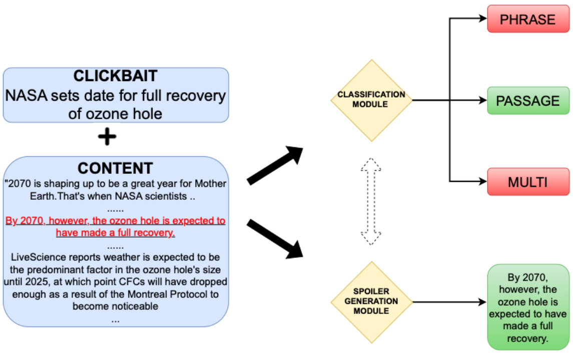

CLICKBAIT
NASA sets date for full recovery of ozone hole
+
CONTENT
"2070 is shaping up to be a great year for Mother Earth.That's when NASA scientists ..
......
By 2070, however, the ozone hole is expected to have made a full recovery.
......
LiveScience reports weather is expected to be the predominant factor in the ozone hole's size until 2025, at which point CFCs will have dropped enough as a result of the Montreal Protocol to become noticeable
...
CLASSIFICATION MODULE
PHRASE
PASSAGE
MULTI
SPOILER GENERATION MODULE
By 2070, however, the ozone hole is expected to have made a full recovery.

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

2 Approach

In this part of the paper, we will explore our strategy for addressing sub-task 1 and provide a concise overview of sub-task 2, highlighting the connection between them.

Our chosen method involves the use of Long Short-Term Memory (LSTM) networks, augmented with an embedding layer for effective data processing and representation learning. We employ bidirectional LSTMs to enhance the model's ability to grasp the context by interpreting sequences in both forward and reverse directions. To facilitate multi-class classification, we implement a softmax activation function to determine the class probabilities.

Data Preprocessing: The first step consists of data preprocessing, which entails tokenizing the text into individual tokens, employing regular expressions (regex) for identifying patterns and cleaning the text, and utilizing vectorization techniques like TF-IDF to convert text into numerical representations. Further preprocessing measures include eliminating stop words, transforming all text to lowercase, and removing contractions to normalize the data.

Hyperparameter Tuning: Adjusting the model's hyperparameters is essential for achieving the best performance. Key parameters consist of the learning rate, which governs how weights are updated during training iterations; batch size, which specifies how many samples are evaluated before updating the model; and details of the architecture like the count of LSTM layers or the number of neurons in each layer.

Validation Metrics: We assess the model's performance using key validation metrics such as accuracy, precision, recall, and the F1 score, all calculated from a confusion matrix. This approach offers a concise yet comprehensive evaluation of the model's predictive capability across various classes.

Loss Function and Optimizer: We use the categorical cross-entropy loss function for its proficiency in multi-class classification. An optimizer minimizes this loss, updating the model's weights during training according to calculated gradients.

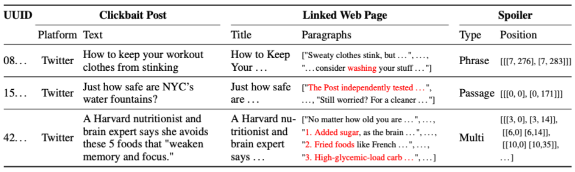
Training: Training involves inputting data into the model, which adjusts its weights to reduce loss. The model's performance is continuously monitored against a validation set to confirm proper learning and facilitate adjustments as needed.

Testing and Evaluation: After training, the model undergoes testing with an unseen test set to evaluate its generalization ability. Its performance is measured using the specified validation metrics, providing a thorough assessment of its real-world applicability.

3 Data

For these tasks, our dataset is the Webis Clickbait Spoiling Corpus 2022, comprising **5,000** posts tagged with spoilers and distributed in a fixed **64/16/20** split for training, validation, and testing across both subtasks. The training and validation datasets were accessible publicly, whereas access to the test dataset was provided upon request.

For a given input post, identified by a UUID, the system must produce an output structured as `{"uuid": "<UUID>", "spoilerType": "<TYPE>"}`, where `<TYPE>` can be either `phrase, passage, or multi` . The requirements for generating a spoiler include the clickbait post, its associated document, and the identified spoiler type. Consequently, the system should output `{"uuid": "<UUID>", "spoiler": "<SPOILER>"}`, with `<SPOILER>` representing the designated spoiler for the clickbait post (Figure 2).



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

Figure 2: Dataset example as they would appear in our corpus(JSONL tabular format)

4 Tools

Listed below are some of the tools which will be used during the implementation.

1. **Numpy**: Crucial for executing numerical computations and managing arrays, matrices, and data across multiple dimensions with high efficiency.

2. **Pandas**: Optimal for editing and analyzing data, particularly adept at managing tabular data seamlessly.

3. **Matplotlib**: Offers an extensive array of plotting functions, critical for the visual exploration and analysis of datasets.

4. **Scikit-learn**: Provides essential tools for data preprocessing, feature extraction, and splitting data into training and testing sets, vital for machine learning endeavors.

5. **NLTK (Natural Language Toolkit)**: An extensive library dedicated to natural language processing (NLP), ideal for analyzing textual data and linguistic studies.

6. **SpaCy**: A sophisticated NLP library, equipped with powerful preprocessing capabilities such as tokenization and part-of-speech tagging.

7. **Gensim**: Focuses on topic modeling and measuring document similarity, invaluable for text vectorization and analysis.

8. **TensorFlow**: Offers a suite of functionalities beyond deep learning, including data augmentation and complex operations, useful for tasks not related to LSTM.

9. **PyTorch**: Comparable to TensorFlow, it shines with its dynamic computation graph and tensor management, applied in contexts beyond LSTM modeling.

10. **Scipy**: Augments scientific computing by offering tools for optimization, linear algebra, and statistical analysis, serving as a complement to Numpy.

5 Learning Outcomes

Successfully completing this project will enhance our comprehension and expertise in LSTMs, hyperparameters, and evaluation metrics. This endeavor has provided us with valuable insights into methodologies adopted by international teams. Additionally, we will endeavor to meet or exceed (hopefully) established baselines, striving for excellence in our implementation. Furthermore, the utilization of tools and libraries outlined in section 4 will enrich our practical understanding of applying theoretical knowledge to real-world situations.

6 Personal Role

In this project, I will mainly contribute in the Evaluation part, more precisely using Validation metrics on the Validation and Test dataset, I will also help my team in choosing the right Hyperparameter settings. I will also contribute in the Programming part of the project. As an Engineer and having a background in Mathematics I will also help my team in enhancing the LSTM model (Algorithm), ultimately improving the model's performance to get a satisfactory accuracy.

References

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)*.

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.

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