# Chick Adams at SemEval-2023 Task 5: Using RoBERTa and DeBERTa to extract post and document-based features for Clickbait Spoiling

**Ronghao Pan**[1], **José Antonio García-Díaz**[1],
**Francisco García-Sánchez**[1], **Rafael Valencia-García**[1]
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan@um.es, joseantonio.garcia8, frgarcia, valencia}@um.es

## Abstract

In this manuscript, we describe the participation of the UMUTeam in SemEval-2023 Task 5, namely, Clickbait Spoiling, a shared task on identifying spoiler type (i.e., a phrase or a passage) and generating short texts that satisfy curiosity induced by a clickbait post, i.e. generating spoilers for the clickbait post. Our participation in Task 1 is based on fine-tuning pre-trained models, which consists in taking a pre-trained model and tuning it to fit the spoiler classification task. Our system has obtained excellent results in Task 1: we outperformed all proposed baselines, being within the Top 10 for most measures. Foremost, we reached Top 3 in F1 score in the *passage spoiler* ranking.

## 1 Introduction

Clickbait is a term used to describe social media posts that aim to attract users' attention and get them to click on those posts or visit a website. This is achieved through catchy titles that often exaggerate or distort the actual content of the post in order to generate traffic or advertising income (Hagen et al., 2022). Clickbait is considered inappropriate because its resolution is often ordinary or trivial, consisting of little more than a sentence, a short passage or a list of things that could easily have been included in the post. For this reason, the Clickbait Spoiling shared task (Fröbe et al., 2023a), proposed at SemEval 2023, consists of counteracting the negative effects of clickbait providing a spoiler or preview that reveals the main content of a post without requiring the user to click through to the website. The task is divided into two sub-tasks:

1. **Spoiler type classification**: The task is to classify the spoiler type that the clickbait post warrants. There are basically three types of spoilers: (1) *phrase spoilers* consisting of a single word or phrase from the linked document, (2) *passage spoilers* consisting of one or a few sentences of the linked document, and (3) *multipart spoilers* consisting of more than one non-consecutive phrases or passages of the linked document.

2. **Spoiler generation**: The task is to generate the spoiler for the clickbait post.

One of the novel features of this shared task is the integration of the TIRA platform for the reproducibility of tasks (Fröbe et al., 2023b), preventing third parties who want to evaluate the state of the art of a task on other datasets from having to reimplement the participants' software (Fröbe et al., 2023).

Advances in deep learning have had a major impact on Natural Language Processing (NLP), significantly improving the ability of computation to understand and process human language. Moreover, the rise of modern Large Language Models (LLMs) based on Transformers (Kalyan et al., 2021) has enabled an increase in the performance of many tasks, such as classification, sentiment analysis, and automatic translations (Bozinovski, 2020). Our contribution is focused on Task 1 (i.e., spoiler-type classification). We propose the use of a transfer learning approach, which consists of fine-tuning already pre-trained Transformer-based models with a large corpus. In this way, a large amount of training data is not needed to achieve better performance in the classification task, since the general knowledge of the pre-trained model is used and tuned for a specific task, as in this case of spoiler-type classification. Our system has obtained really high results in Task 1, reaching the Top 10 in most of the measures and even the Top 3 in the F1 score in the *passage spoiler* ranking.

The remainder of this paper is organized as follows. Section 2 provides a summary of important details about the task setup. Section 3 offers an overview of the proposed system for the first subtask. The experimental setup is described in

*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 624–628
July 13-14, 2023 ©2023 Association for Computational Linguistics

Section 4. Section 5 discusses the results of the experiments, and finally, the conclusions are put forward in Section 6.

## 2 Background

We made use exclusively of the dataset provided by the organizers, two pre-trained models, post-based features, and document-based features.

The dataset consisted of 5000 posts written in English and distributed into three subsets, namely, training, validation, and testing, with 3200, 800, and 1000 posts, respectively. Most spoiled clickbait posts comes from Twitter (47.5O%) and Reddit (36%), whereas the Facebook account contributes a bit less (16.5%). Our first experiment consisted in examining the number of examples of each type (either "phrase", "passage", "multi") to see if the dataset was balanced or not. The statistics of the dataset concerning the first subtask are shown in Table 1. There is a significant imbalance between labels, especially the multipart type since in the training set, the number of spoiler instances of phrase and passage type exceeds the multipart type by more than 50%.

Table 1: Corpus statics for Task 1

| Split | Phrase | Passage | Multipart | Total |
|---|---|---|---|---|
| Train | 1367 | 1274 | 559 | 3200 |
| Dev | 335 | 322 | 143 | 800 |
| Test | - | - | - | 1000 |

The data comes in JSON line format (.jsonl), where each line contains a clickbait post and the manually cleaned version of the linked document. Therefore, to predict the type of spoiler, we have information about the clickbait post, such as the post text and target title, and information about the main content of the linked web page in the format of several paragraphs of different lengths.

## 3 System Overview

For solving Task 1, we built the system whose architecture is depicted in Figure 1. In a nutshell, our system works as follows. First, the dataset was preprocessed using the techniques pointed out in Section 3.1. Second, we extracted the post and document features of each item in the dataset. Third, we fine-tuned different pre-trained models for the classification of spoilers types. Next, different ensemble learning techniques were evaluated to see if the post and document features complemented each other and improved the model performance. Finally, the best performing model was chosen to be tested with the test split. The source code is available at Github[1]

### 3.1 Preprocessing stage

Our preprocessing stage consists in the following procedure to clean both clickbait related data and linked documents:

1. To remove social media language, such as hyperlinks.
2. To replace all hashtags and mentions with *#[HASHTAG]* and *@[USER]*.
3. To remove all email links.
4. Clickbaits on social networks require a lot of clean-ups, but it is inefficient to clean up each item, so a general clean-up approach was applied in this case.
    - To remove all emojis.
    - To expand some contractions.

Next, after cleaning up the texts, the clickbait posts have been divided into two parts: post-based features and document-based features. The post-based feature is formed by the union of the post and the target title. When analyzing the texts of the linked documents, we detected that they are formed by paragraphs of different lengths, with some of them being very long texts, which makes it difficult to learn the models. Therefore, to reduce the size of long paragraphs, a Sentence-Transformer model (Reimers and Gurevych, 2020) called *sentence-transformers/all-MiniLM-L12-v2*[2] have been used to extract only the part that is more closely related to the text of the post. Once all the paragraphs do not exceed 200 tokens, we selected the one that is most related to the text of the post by leveraging the same Sentence-Transformer model.

### 3.2 Classification model

Our approach for Task 1 is based on fine-tuning different pre-trained models, namely, RoBERTa and DeBERTa. The details are provided in the next subsections.

[1] https://github.com/ronghaopan/Semeval-Task5-Clickbait

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

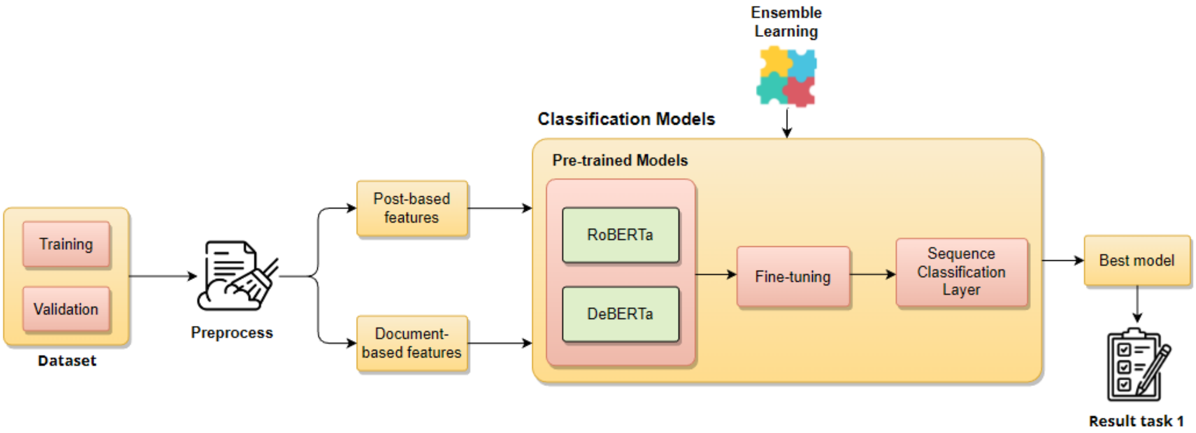

Figure 1: Overall system architecture.

### 3.2.1 Pre-trained models

The pre-trained models used are as follows:

- **RoBERTa-large**[3]: This is a variant of the RoBERTa language model developed by Facebook AI Research (FAIR) in 2019. It is a large-scale pre-trained Transformer-based model, which has 355 million parameters, making it significantly larger than the base version of RoBERTa (125 million parameters) and BERT-large (340 million parameters). Its larger size allows it to capture more complex patterns in the text data and to provide more accurate predictions for a wide range of NLP tasks (Liu et al., 2019).
- **DeBERTa-v2-xlarge**[4]: This is a variant of the DeBERTa (Decoding-enhanced BERT with Disentangled Attention) language model developed by Microsoft Research Asia in 2021. It is a large-scale pre-trained transformer-based model, which has 3.3 billion parameters, making it significantly larger than the base version of DeBERTa (345 million parameters) (He et al., 2021).

### 3.2.2 Fine-tuning

Fine-tuning is a deep learning model training process in which a pre-trained model is taken as a basis and additionally trained on a specific task with a smaller dataset. In this way, the prior knowledge of the pre-trained model can be exploited and its behavior adjusted. In this case, in the absence of a large enough amount of training data, as shown in Table 1, this process improves the overall performance of the model compared to training it from scratch with the corpus provided by the organizers.

## 4 Experimental Setup

After the preprocessing stage, we have two datasets for Task 1, one with post features and one with document features (see Section 3.1). Both datasets have 3200 items for training and 800 for evaluation. The hyperparameters used for fine-tuning the RoBERTa-large and DeBERTa-v2-xlarge are 10 epochs, 0.01 in weight decay, a batch training of 8, and a learning rate of 1e-5.

Having two datasets with different features, different ensemble learning strategies have been evaluated on the trained models to see if the two features complement each other and improve the overall performance of the model. In ensemble learning, the output of each trained model with a set of textual features is combined by averaging the prediction (*mean*) or choosing the prediction with higher probability (*max*).

## 5 Results

Each model was evaluated using the validation split (see Section 2). The result for Task 1 is depicted in Table 2. As it can be observed, the performance of the models trained with post-based features is much better than that of the models trained with document-based features. In particular, it is a 14.12% better in the case of RoBERTa, and a 21.01% better in the case of the DeBERTa model. With respect to the pre-trained models, the fine-tuned DeBERTa-v2 performs better in the post-based training set, and the fine-tuned RoBERTa-large performs better in the document-based set.

[3] https://huggingface.co/roberta-large

[4] https://huggingface.co/microsoft/deberta-v2-xlarge

From Table 2), it can be also noticed that *mean*-based and *max*-based ensemble learning attempts were not able to improve the results obtained with post features alone and it can be concluded that the linked documents do not complement the target title and post text of a clickbait post.

| **Model** | **M-F1** | **W-F1** |
|---|---|---|
| **Post-based features** | | |
| RoBERTa-large | 0.7157 | 0.7241 |
| DeBERTa-v2 | 0.7156 | **0.7259** |
| **Document-based features** | | |
| RoBERTa-large | 0.5773 | **0.5829** |
| DeBERTa-v2 | 0.4977 | 0.5158 |
| **Ensemble learning** | | |
| Mean | 0.6088 | 0.6159 |
| Max | 0.6339 | 0.6412 |

Table 2: Results for Task 1 with the validation split, reporting the macro F1-score (M-F1) and the weighted F1-score (W-F1).

To evaluate the performance of the best model for Task 1 (DeBERTa-v2 fine-tuned with post-based features) and to find out in which cases the model gives wrong predictions, a confusion matrix has been used. The confusion matrix is a table showing the number of correct and incorrect predictions made by a classification model on a dataset. The confusion matrix is shown in Figure 2. In particular, taking into account this confusion matrix, it can be observed that the model labeled wrongly a 7.25% of *passage* as *multipart* spoilers and 8.75% *multipart* as *passage* spoilers.

Our approach has obtained excellent results in Task 1, reaching to be inside Top 10 for most measures. In particular, we reached Top 3 in F1 score in the *passage spoiler* ranking.

### 5.1 Post-evaluation

During the post-evaluation phase, we evaluated the model performance using a combination of post-based and document-based features as the training set. We evaluated the DeBERTa-large and RoBERTa-large models, and the results are shown in Table 3. The combined features approach has shown an improvement in the overall performance with respect to the ensemble learning approach and fine-tuning models using single features. DeBERTa-large is the most effective model, achieving a macro F1-score of 73.48% and a weighted F1-score of 74.15%, outperforming the best model of the fine-tuning approach by 1.563% and the best model of the ensemble learning approach by 10.033% in terms of weighted F1-score.

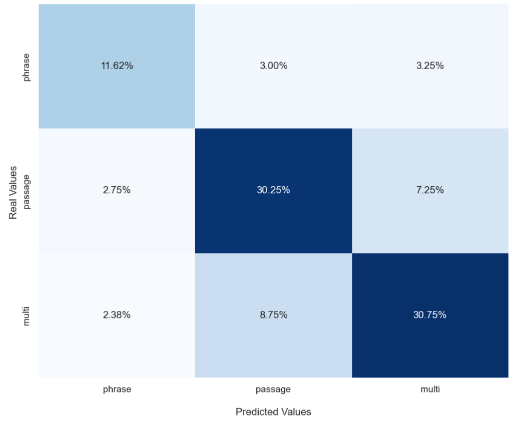

Figure 2: Confusion matrix of the DeBERTa-v2 fine-tuned with post-based features.

Table 3: Results for Task 1 with the validation split using combined features as training dataset, reporting the macro F1-score (M-F1) and the weighted F1-score (W-F1).

| **Model** | **M-F1** | **W-F1** |
|---|---|---|
| RoBERTa-large | 0.72660 | 0.73301 |
| DeBERTa-large | 0.73477 | **0.74153** |

## 6 Conclusion

In this working notes we describe the participation of the UMUTeam in the shared task 5 of SemEval 2023, namely, Clickbait Spoiling. In this shared task, the participants were required to identity spoiler type (i.e., a phrase or a passage) and generating short texts that satisfy curiosity induced by a post clickbait, i.e. generating spoilers for the clickbait post. Our system has obtained excellent results in Task 1, outperforming all proposed baselines, achieving the Top 10 for most measures and a Top 3 F1-score in the *passage spoiler* ranking.

As further work, we are planing to measure the correlation between headlines related to clickbaits concerning the financial domain in Spanish. For this, we will extend the dataset published in (García-Díaz et al., 2023) by incorporating news from financial news sites coming from Spanishspeaking countries and will apply the Spanish and multilingual variants of RoBERTa and DeBERTa. Besides, we intend to extend this dataset with the inclusion of content from satirical media (García-Díaz and Valencia-García, 2022) and measure the impact of clickbait in datasets concerning political ideology (García-Díaz et al., 2022). On the other hand, as commented in Section 5, the most frequent error in our model is that it confuses *passage*-type spoilers with *multiple*-type spoilers. Therefore, as future work, we plan to use volumetric and lexical diversity features in the training set to mitigate this error.

## 7 Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

## References

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)*, 44(3).

Maik Fröbe, Tim Gollub, Benno Stein, Matthias Hagen, and Martin Potthast. 2023a. SemEval-2023 Task 5: Clickbait Spoiling. In *17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023b. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous integration for reproducible shared tasks with tira.io. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, page 236–241, Berlin, Heidelberg. Springer-Verlag.

José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023. Smart analysis of economics sentiment in spanish based on linguistic features and transformers. *IEEE Access*, 11:14211–14224.

José Antonio García-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín Valdivia, Francisco García-Sánchez, L Alfonso Ureña-López, and Rafael Valencia-García. 2022. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguaje Natural*, 69:265–272.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. 2022. Clickbait spoiling via question answering and passage retrieval. pages 7025–7036.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. AMMUS : A survey of transformer-based pretrained models in natural language processing. *CoRR*, abs/2108.05542.

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. *CoRR*, abs/1907.11692.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.