Clickbait Detection
POC Report for NLP Course, Winter 2025

Jakub Sawicki
Warsaw University of Technology
jakub.sawicki3.stud@pw.edu.pl

Jędrzej Sokołowski
Warsaw University of Technology
jedrzej.sokolowski.stud@pw.edu.pl

Wiktor Woźniak
Warsaw University of Technology
wiktor.wozniak2.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Clickbait headlines are designed to grab attention by leaving out key information or sparking curiosity. It makes them tricky to detect automatically. This milestone examines the latest methods and resources for identifying clickbait, reviews current datasets, and proposes a new approach. The method combines publicly available datasets with pretrained transformer-based models to deliver reliable and generalizable results.

1 Introduction

Clickbait refers to headlines or social media posts that are crafted to get people to click, often by teasing curiosity, using sensational language, or leaving out important information. Usually, the content doesn't fully deliver what was promised in the headline. This is common across news sites and social media platforms. (Setlur, 2018; Omidvar et al., 2018). As a result, it is difficult to determine whether the headlines are genuinely grabbing our attention or if they are manipulative. As a result, it is more challenging for readers to assess the quality of information.

Detecting clickbait automatically is difficult. The difference between a catchy headline and a misleading one can be subtle. Even humans often disagree on what counts as clickbait. Effective detection systems need to look at both the language of the headline and how it relates to the actual article, because a headline may seem misleading without the full context (Omidvar et al., 2018). Importantly, clickbait exists on a spectrum. This makes the problem more complex and motivates research into models that can predict degrees of "clickbaitiness" rather than just labeling content as clickbait or not.

2 Background and Motivation

Efforts to automate clickbait detection have evolved from simple hand-crafted rules to sophisticated machine learning and deep learning models. The Clickbait Challenge 2017 advanced the field by introducing the idea of treating clickbait detection as a regression problem by rating headlines along a continuous scale (Potthast et al., 2018a). This has encouraged the development of more elaborate systems that go beyond simple binary categorization, capturing the subtlety and diversity of misleading headline techniques.

Beyond detection, clickbait "spoiling" has emerged as an additional task, where systems not only label clickbait but also generate or extract the missing information that these headlines withhold (Hagen et al., 2022). This approach recognizes that clickbaits often lack simple facts, and offers the withheld information directly to users reduces unnecessary clicks and increases transparency.

As clickbait tactics became more sophisticated and sometimes automated through AI, the need for robust and adaptable models grew. Advances in the field improved due to the growing availability of open datasets and shared task competitions, allowing for creation of reproducible benchmarks and faster scientific progress.

3 Related works

3.1 Feature Engineering and Hand-crafted Methods

Early clickbait detection systems relied heavily on feature engineering. These models examined a wide range of linguistic, syntactic, and content-based data taken from headlines and their corresponding articles (Elyashar et al., 2017; Zuhroh and Rakhmawati, 2019). Typical features included word and character counts, punctuation use, n-grams, informal language markers, and the de-

gree to which the headline matched the opening paragraphs of the article. Some more advanced approaches also incorporated text extracted from images using OCR, along with metadata such as posting time or how long a post was visible (Elyashar et al., 2017). Tools like Stanford CoreNLP provide additional syntactic features, including contraction frequency and dependency lengths (Chakraborty et al., 2016). With carefully selected features and preprocessing steps such as lowercasing and removing stopwords, these systems provided a solid background for the field.

3.2 Model Architectures: Classical, Neural, and Transformer-based

Classical machine learning models, such as Random Forests and SVMs, perform well when combined with well-designed features, reaching accuracy levels of up to 93% in some studies (Chakraborty et al., 2016). As deep learning became more common, the field shifted away from manual feature engineering. Models that represented headlines using Word2Vec, GloVe, and similar embeddings allowed CNNs and RNNs to automatically learn deeper patterns in the data (Wongsap et al., 2018; Zheng et al., 2018). Architectures such as dedicated CNNs and MLPs outperformed traditional methods, especially on larger or multilingual datasets. They were better at capturing typical clickbait techniques.

More recently, transformer-based models such as DeBERTa and RoBERTa have set a new standard. These models are fine-tuned not only for clickbait detection but also for the newer task of clickbait spoiling. Because transformers are pretrained on massive text data using objectives like masked language modeling and question answering, they achieve state-of-the-art results for both classification and spoiler generation.

4 Open Benchmark Datasets

Four major open datasets have played a key role in clickbait detection research. Each of them offers a different perspective on how clickbait appears across platforms and languages.

Webis Clickbait Spoiling Corpus 2022 (Fröbe et al., 2023) is a comprehensive dataset designed for both clickbait detection and spoiling. It contains 5,000 manually annotated posts collected from Twitter, Reddit, Facebook, and other social platforms. Not only were the headlines labeled as clickbait or not, but a short piece of text was also written that reveals the information that is missing in a headline (Hagen et al., 2022).

Wikinews Clickbait Corpus includes 18,513 news articles from Wikinews, with 7,623 labeled as clickbait by crowdsourced workers (Chakraborty et al., 2016). To create a balanced dataset, 7,500 non-clickbait articles were randomly selected. Each article was judged by at least three independent experts, and labels were assigned based on majority vote, improving consistency and reliability.

Thai Headline News Dataset expands clickbait research beyond English by providing 5,000 Thai news headlines, each annotated by two experts (Wongsap et al., 2018). For benchmarking, the dataset keeps 2,000 headlines labeled as clickbait and 2,000 as non-clickbait.

Chinese News Headlines Dataset contains 14,922 headlines from major Chinese news portals (Zheng et al., 2018). Half of the headlines are labeled as clickbait. The dataset also includes metadata about article types. Its size and diversity make it valuable for multilingual and cross-domain clickbait detection, particularly when evaluating advanced neural models.

These datasets enable robust model development, fair benchmarking, and strong generalization studies across both detection and spoiling tasks.

5 Proposed Solution and Project Plan

5.1 Objective

This project aims to develop a practical and interpretable clickbait detection system for headlines using publicly available benchmark datasets. The main research questions guiding this work are:

- How effectively can classical machine learning models perform on clickbait detection using hand-crafted linguistic and content-based features?
- To what extent do embedding-based neural models improve performance over classical approaches on these datasets?
- Can transformer-based models, fine-tuned on clickbait datasets, provide superior accuracy and generalization while maintaining efficiency?

5.2 Feature Extraction

The proposed system will extract a comprehensive set of features from headlines and their associated articles, including but not limited to:

- N-gram frequencies (unigrams, bigrams) in headlines.
- Headline and article length (character and word counts).
- Overlap of headline terms with article opening sentences.
- Presence of informal words and pronouns.
- Sentiment score of the headline.
- Counts of punctuation and exclamatory words.

Standard NLP libraries such as NLTK or spaCy will be used to automate preprocessing and feature generation.

5.3 Models to be Evaluated

We plan to implement and compare the following models:

- **Classical Machine Learning Models:** Logistic Regression, Support Vector Machines (SVM), Random Forests, and Gradient Boosting Trees trained on the hand-crafted features.
- **Neural Embedding Models:** Feedforward Neural Networks and Convolutional Neural Networks (CNNs) leveraging pre-trained word embeddings (Word2Vec, GloVe).
- **Transformer Models:** Fine-tuned transformer-based architectures such as BERT and RoBERTa pre-trained on related NLP tasks and adapted for clickbait detection.

5.4 Datasets

We will use the Webis Clickbait Spoiling Corpus from different years as the primary data source. Data will be split into training and testing sets, ensuring balanced representation of clickbait and non-clickbait examples.

5.5 Evaluation

Performance will be measured using standard metrics:

- Accuracy and F1 score for classification.
- Mean squared error for measuring clickbait strength.

Additionally, feature importance analysis will be provided to identify which textual patterns contribute most to the detection task.

6 Proof of Concept

6.1 Data

We used data from **Webis Clickbait Corpus 2017** (Potthast et al., 2018b). It consists of Twitter posts linking to news articles. The titles of articles are available with the degree of clickbaitiness in $[0, 1]$ interval. Larger values indicate that the observation is a clickbait.

As our datasets we used **clickbait17-train-170331** and **clickbait17-train-170630** folders available for download. To combat the almost 3 to 1 class imbalance, we extended our data using **Webis Clickbait Spoiling Corpus 2022** (Fröbe et al., 2023) dataset only consisting of clickbait entries.

From each item, the post text was standardized by flattening the text field into a single string, producing a normalized headline representation. The labels were binarized using a threshold of 0.5. Entries with missing or empty textual content were removed. The resulting cleaned dataset was then stratified into training, validation and testing sets (60/20/20) preserving class balance in both sets.

54 rows were removed because of missing or empty headlines. Finally, we obtained 25943 observations, 10377 of them were assigned to the training set and 7783 to the validation and test set each.

6.2 Data processing

We implemented a custom feature transformer to capture simple textual features. The extractor computes a fixed set of interpretable features for each headline. For every input text, the following attributes are derived:

- **Character count:** Total number of characters in the headline.

- **Word count:** Total number of whitespace-separated tokens.
- **Punctuation cues:** Counts of exclamation marks and question marks.
- **All-caps indicator:** Whether the entire headline is written in uppercase.
- **Sentiment polarity and subjectivity:** Obtained via `TextBlob`, capturing coarse affective tendencies.
- **Interrogative start:** A binary flag indicating whether the headline begins with typical question words (*who, what, where, why, how*), a pattern frequently associated with curiosity-inducing clickbait.

The obtained features are then added to the datasets.

6.3 Exploratory Data Analysis

By combining two datasets, class balance was improved, their distribution is shown in Figure 1.

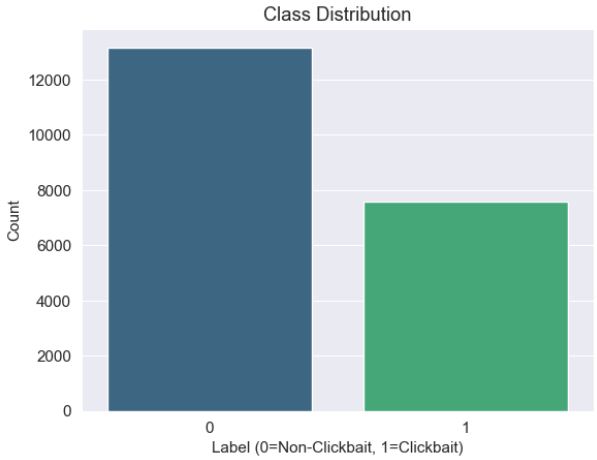

Figure 1: Class Distribution

To better understand the lexical characteristics of clickbait and non-clickbait headlines, we performed a frequency analysis of words in the training set. Headlines were first tokenized using `NLTK`'s word tokenizer, converted to lowercase, and filtered to remove stopwords and non-alphanumeric tokens.

For each class (*clickbait* and *non-clickbait*), the top 10 most frequent words were identified. The results are shown on Figures 2 and 3.

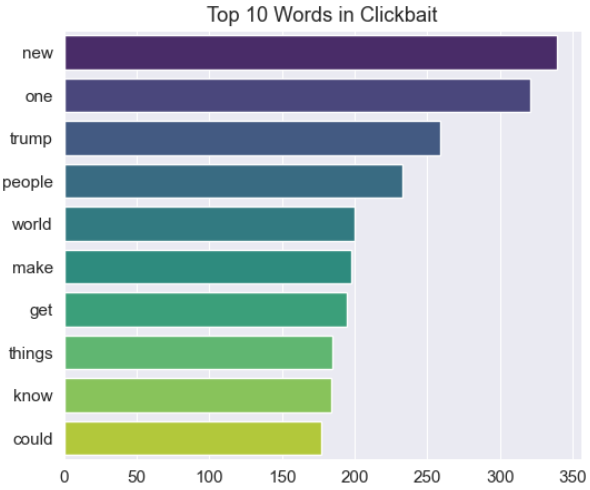

Figure 2: Top 10 words in clickbait headlines from the training set

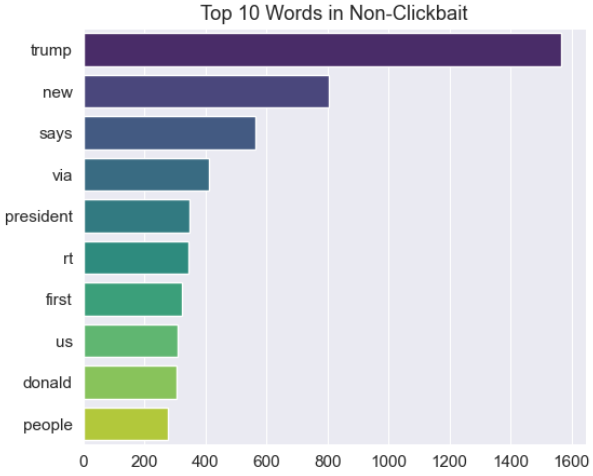

Figure 3: Top 10 words in non-clickbait headlines from the training set

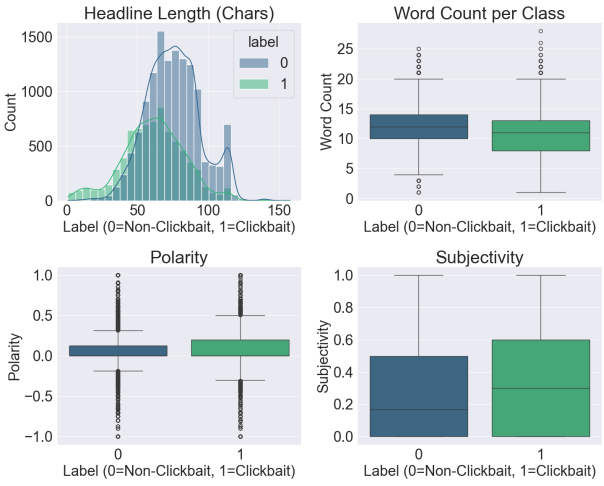

Figure 4: Distribution of selected features within clickbait and non-clickbait groups

We conducted further analysis to examine headline length, word count, and distributions of hand-crafted features such as punctuation counts, sentiment, and interrogative starts. Plots highlighting differences between clickbait and non-clickbait headlines are shown on Figures 4 and 5. In Figure 4 we can see subtle differences between the two labels, where clickbait headlines tend to be slightly shorter, more polarized and subjective. Figure 5 demonstrates that the proportion of headlines involving exclamation and question marks, as well as the usage of all caps case is vastly greater in clickbait ones. Similarly, they begin with question words much more often compared to non-clickbait.

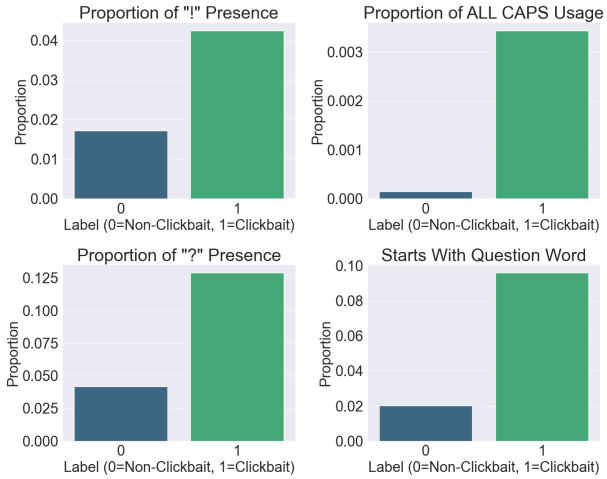

Figure 5: Distribution of selected features within clickbait and non-clickbait groups

6.4 Models

We decided to verify the results of classification for three approaches. We used Random Forest as a classical machine learning approach, a Feedforward Neural Network and a Transformer model - DistilBERT.
We used training set for training and validation set for hyperparameters tuning. For now, we report the obtained results on the validation set.

6.5 Classical Machine Learning Model

We implemented a classical supervised model to classify headlines as clickbait or non-clickbait. The input features combined both textual and hand-crafted cues using a `FeatureUnion`:

- **TF-IDF n-grams:** Unigrams and bigrams extracted from headline text, limited to 1000 features.
- **Hand-crafted features:** Stylistic and lexical features such as punctuation counts, sentiment scores, all-caps usage, and interrogative start.

These features were fed into a `RandomForestClassifier`. We performed cross-validation GridSearch for finding the best hyperparameters. We tested different number of trees, maximal depth and maximal number of features. The optimal hyperparameters identified during training were a maximum tree depth of None, sqrt feature selection, and 200 estimators. We present the results on test set in Table 1.

Table 1: Random Forest Model Performance (Validation Set)

Class	Prec.	Rec.	F1	Supp.
0 (Non-Clickbait)	0.82	0.87	0.85	4939
1 (Clickbait)	0.75	0.66	0.71	2844
Accuracy		0.80		7783
Macro Avg	0.79	0.77	0.78	7783
Weighted Avg	0.79	0.80	0.79	7783

We provide also a confusion matrix computed on the test set in Figure 6.

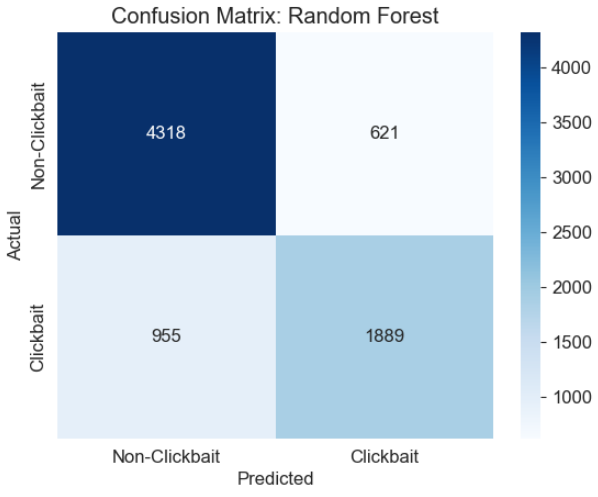

Figure 6: Confusion matrix of Random Forest.

6.5.1 Feedforward Neural Network

Headlines were represented using pre-trained 50-dimensional GloVe embeddings, where each headline vector was computed as the average of its constituent word embeddings. These fixed-size representations were used as input to a feed-forward neural network consisting of three hidden layers

(128, 64, and 32 units) with ReLU activations and dropout regularization. The model was trained for clickbait binary classification using PyTorch.

Adam optimizer with 0.001 learning rate and cross-entropy loss were used. The model was trained from scratch for 100 epochs with 32 batch size.

The validation loss was monitored every epoch and the best performing model on validation set was saved. The results of the best model can be found in Table 2 and Figure 7.

Table 2: Feed-Forward Neural Network Performance (Validation Set)

Class	Prec.	Rec.	F1	Supp.
0 (Non-Clickbait)	0.79	0.86	0.82	4939
1 (Clickbait)	0.71	0.59	0.65	2844
Accuracy		0.76		7783
Macro Avg	0.75	0.73	0.74	7783
Weighted Avg	0.76	0.76	0.76	7783

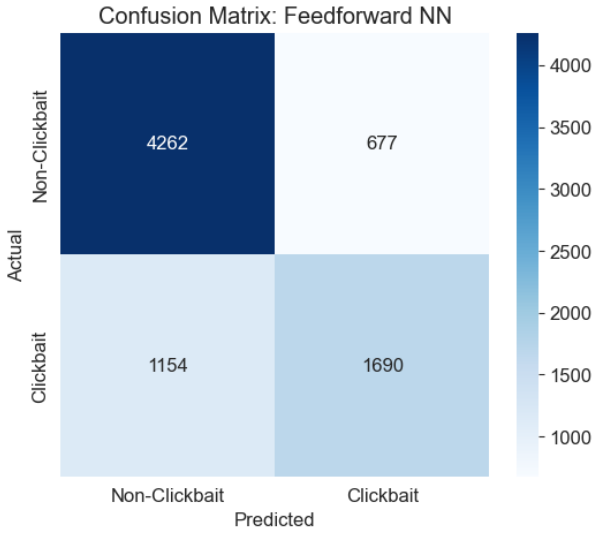

Figure 7: Confusion matrix of Feedforward NN model.

6.5.2 Transformer-based approach

To leverage contextual embeddings, we fine-tuned a DistilBERT model for binary clickbait classification. The text data was tokenized using the `DistilBertTokenizer` with truncation and padding to a maximum sequence length of 64 tokens. Fine-tuning was performed using the Hugging Face `Trainer` API with 2 epochs and a batch size of 8. Longer training increased the validation loss. The validation loss was monitored every 100 steps. The best model in terms of it was saved.

The results are shown in Table 3 and Figure 8.

Table 3: DistilBERT Model Performance (Validation Set)

Class	Prec.	Rec.	F1	Supp.
0 (Non-Clickbait)	0.87	0.86	0.87	4939
1 (Clickbait)	0.77	0.78	0.77	2844
Accuracy		0.83		7783
Macro Avg	0.82	0.82	0.82	7783
Weighted Avg	0.83	0.83	0.83	7783

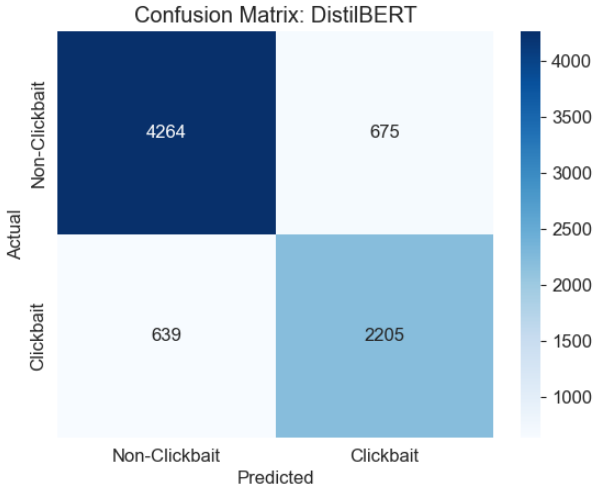

Figure 8: Confusion matrix of DistilBERT model.

6.6 Clickbaitness Analysis

For the POC part, we conducted a clickbaitness analysis using the measure proposed by (Urban et al., 2025). Figure 9 presents the histogram of clickbaitness scores for the test dataset. As observed, most posts have clickbaitness values between 0.3 and 0.5. This distribution aligns with the label distribution shown in Figure 1, which indicates a higher number of non-clickbait posts compared to clickbait posts.

The analysis also highlights that this task is challenging due to the lack of extreme clickbaitness values, suggesting that most posts fall into a moderate range rather than being clearly clickbait or non-clickbait.

For the final report, we plan to extend this analysis by comparing the clickbaitness measure with the probabilities predicted by our models. Additionally, we aim to model post clickbaitness based on model predictions and provide deeper insights

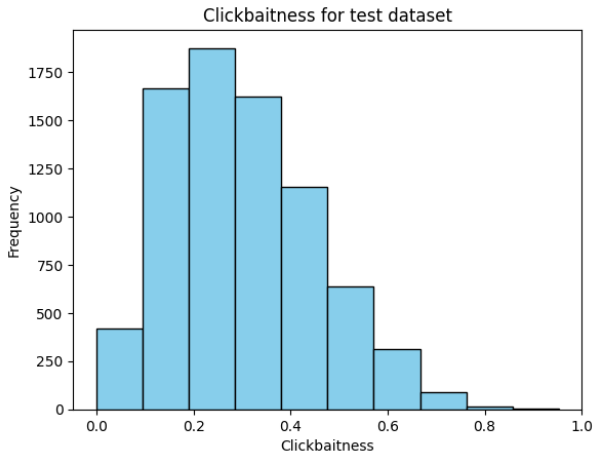

Figure 9: Histogram of clickbaitness scores in the test dataset.

into the relationship between predicted probabilities and the clickbaitness.

7 Summary

This project aims to build a clear and effective system to detect clickbait headlines. We will compare different approaches, from traditional machine learning using hand-picked features to advanced neural networks and transformer models. By using benchmark datasets, we will measure how well each method works and also see which headline patterns are most important for detection. The goal is to create a system that is both accurate and easy to understand.

The preliminary results suggest that the most prominent approach from the tested ones is Transformer-based. It demonstrated superior precision–recall balance across both classes.

The NN model achieved comparative results to the RF. RF performed even slightly better on every investigated metric.

Nevertheless, the discrepancy in models' performance was not huge. All three models achieved promising results and were able to effectively handle the minority class.

References

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. 10.

Aviad Elyashar, Jorge Bendahan, and Rami Puzis. 2017. Detecting clickbait in online social media: You won't believe how we did it.

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, Toronto, Canada, July. Association for Computational Linguistics.

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

Amin Omidvar, Hui Jiang, and Aijun An. 2018. Using neural network for identifying clickbaits in online news media.

Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. 2018a. The clickbait challenge 2017: Towards a regression model for clickbait strength.

Martin Potthast, Tim Gollub, Matti Wiegmann, Benno Stein, Matthias Hagen, Kelsey Komlossy, Sebastian Schuster, and Eduardo Pedroza Garcia Fernandez. 2018b. Webis clickbait corpus 2017 (webis-clickbait-17). Data set.

Amrith Rajagopal Setlur. 2018. Semi-supervised confidence network aided gated attention based recurrent neural network for clickbait detection.

Tymoteusz Urban, Mateusz Kubita, and Wojciech Michaluk. 2025. Clickbait news detection and analysis.

Natnicha Wongsap, Tastanya Prapphan, Lisha Lou, Sarawoot Kongyoung, Sasiwimol Jumun, and Natsuda Kaothanthong. 2018. Thai clickbait headline news classification and its characteristic. In *2018 International Conference on Embedded Systems and Intelligent Technology International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*, pages 1–6.

Hai-Tao Zheng, Jin-Yuan Chen, Xin Yao, Arun Kumar, Yong Jiang, and Cong-Zhi Zhao. 2018. Clickbait convolutional neural network. *Symmetry*, 10:138, 05.

Nurrida Zuhroh and Nur Rakhmawati. 2019. Clickbait detection: A literature review of the methods used. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 6:1, 10.