“Why is this misleading?”: Detecting News Headline Hallucinations with Explanations

Jiaming Shen
Google Research
jmshen@google.com

Jialu Liu
Google Research
jialu@google.com

Dan Finnie
Google
danfinnie@google.com

Negar Rahmati
Google
negarr@google.com

Michael Bendersky
Google Research
bemike@google.com

Marc Najork
Google Research
najork@google.com

ABSTRACT

Automatic headline generation enables users to comprehend ongoing news events promptly and has recently become an important task in web mining and natural language processing. With the growing need for news headline generation, we argue that the *hallucination issue*, namely the generated headlines being not supported by the original news stories, is a critical challenge for the deployment of this feature in web-scale systems Meanwhile, due to the infrequency of hallucination cases and the requirement of careful reading for raters to reach the correct consensus, it is difficult to acquire a large dataset for training a model to detect such hallucinations through human curation. In this work, we present a new framework named ExHalder to address this challenge for headline hallucination detection. ExHalder adapts the knowledge from public natural language inference datasets into the news domain and learns to generate natural language sentences to explain the hallucination detection results. To evaluate the model performance, we carefully collect a dataset with more than six thousand labeled ⟨article, headline⟩ pairs. Extensive experiments on this dataset and another six public ones demonstrate that ExHalder can identify hallucinated headlines accurately and justifies its predictions with human-readable natural language explanations.

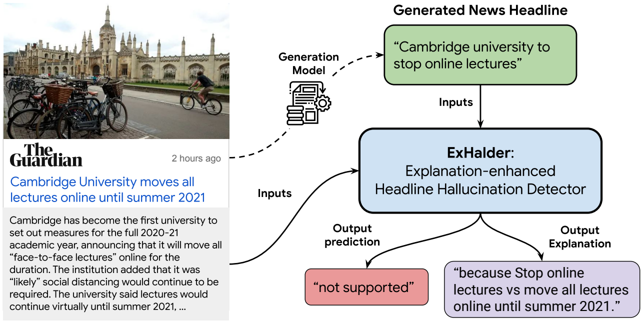
CCS CONCEPTS

- **Computing methodologies** → **Natural language processing**; • **Information systems** → **Web applications**.

KEYWORDS

Hallucination Detection, Natural Language Explanation

ACM Reference Format:
Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Michael Bendersky, and Marc Najork. 2023. “Why is this misleading?”: Detecting News Headline Hallucinations with Explanations. In *Proceedings of the ACM Web Conference 2023 (WWW ’23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583375



Generated News Headline
“Cambridge university to stop online lectures”
Generation Model
Inputs
The Guardian
2 hours ago
Cambridge University moves all lectures online until summer 2021
Inputs
ExHalder: Explanation-enhanced Headline Hallucination Detector
Output prediction
Output Explanation
“not supported”
“because Stop online lectures vs move all lectures online until summer 2021.”

Cambridge has become the first university to set out measures for the full 2020-21 academic year, announcing that it will move all “face-to-face lectures” online for the duration. The institution added that it was “likely” social distancing would continue to be required. The university said lectures would continue virtually until summer 2021, ...

Figure 1: An illustrative example of automated news headline hallucination detection with a model generated natural language explanation.

1 INTRODUCTION

With tens of millions of news articles published every day on the web [9], people are inundated with massive news contents and find them hard to digest. To facilitate more efficient and user-friendly news content consumption, recent works in the industry propose to generate headlines from either a single news article [2] or a set of news articles related to the same event [23]. The generated news headline is intended to serve as a succinct, informative, and accurate summary of its underlying news article(s), and thus it helps the users to quickly grasp the gist of a news story.

To obtain high-quality news headlines, early studies [26, 31, 39] propose *extractive* methods to first extract words from the article title and then organize those salient words into the output headline. More recently, with the advances of natural language generation research [47, 54, 64], more *abstractive* methods are developed to directly summarize the news article into a concise news headline [2, 6, 23, 33, 53]. These abstractive summarization methods typically adopt the encoder-decoder architecture [12, 52] where the encoder synthesizes the knowledge in the news article using vector representations and the decoder outputs the generated headline in a word-by-word fashion. Although overall quality improvements have been made by this approach, people observe that these generation models often will output hallucinated headlines that are not supported by the underlying news articles. For example, in Figure 1, the generation model outputs the headline “*Cambridge university to stop online lectures*” based on the article with the title “Cambridge University moves all lectures online until summer

WWW ’23, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583375

2021". The generated headline is misleading because it suggests that Cambridge University will stop online lectures instead of moving some face-to-face lectures online until the summer of 2021.

In this paper, we study the *news headline hallucination detection* task: given a pair of ⟨news article, news headline⟩, we aim to algorithmically determine if the headline is supported by the underlying article and thus is not misleading. Figure 1 shows an example where the news article indicates Cambridge University will move in-person lectures online for a period of time but the generated news headline suggests the opposite. Therefore, this is a misleading headline and the hallucination detector should predict this headline as "not supported". An intuitive approach to this task is to train a classifier using a large set of ⟨article, headline⟩ pairs with their hallucination labels. However, as those hallucination cases appear infrequently and require deep reading comprehension, such a labeled dataset is usually of small scale and thus forbids us from learning a powerful model that can capture the subtle semantic differences between news articles and news headlines.

To tackle the lack-of-supervision challenge, we propose a novel framework named **ExHalder**, standing for "**Ex**planation-enhanced Headline **Hal**lucination **de**tecto**r**". ExHalder is developed based on two key ideas. First, we observe that there exist many similarities between the headline hallucination detection (HHD) task and the natural language inference (NLI) [5, 35] task. For example, both of them aim to detect if one piece of text ("headline" in the HHD task and "hypothesis" in the NLI task) is supported/entailed by another piece of text ("article" in the HHD task and "premise" in the NLI task). Based on this observation, we propose to pretrain ExHalder using public large-scale NLI datasets [5, 7, 63] and transfer the knowledge learned from the NLI task to the headline hallucination detection task. Second, as the framework name suggests, we propose to go beyond the binary class label and utilize natural language explanations to augment the model learning process. These explanations are particularly useful in the low resource setting (i.e., with limited training data) and help models to generalize better. We demonstrate that the learned ExHalder can generate high-quality human-readable explanations to justify its prediction results. Take the case in Figure 1 for example, ExHalder not only predicts the headline is "not supported" by the news article but also justifies the output with an explanation "because Stop online lectures vs move all lectures online until summer 2021".

To make the best use of these explanations, ExHalder includes three key components: (1) a *reasoning classifier* which receives as input the ⟨article, headline⟩ pair and outputs the class label along with the label explanation, (2) a *hinted classifier* which receives as input the ⟨article, headline, explanation⟩ triplet and predicts the class label, and (3) an *explainer* that generates the natural language explanation based on the input ⟨article, headline⟩ with its known class label. These three components utilize the explanation signals from different angles and work collaboratively within our ExHalder framework. Specifically, during the training phase, we will train the explainer to generate more explanations and use them to augment the original training set for learning the reasoning classifier and the hinted classifier. At the inference stage, we first input the test ⟨article, headline⟩ tuple into the reasoning classifier to obtain its predicted class and generated explanation. Then, we concatenate the explanation with the input tuple and feed them together into the hinted classifier to obtain another class prediction. Finally, we aggregate these two predictions and return the final predicted class with its corresponding explanation.

We test the effectiveness of the ExHalder framework on seven hallucination detection datasets from different domains. Our results demonstrate that ExHalder achieves state-of-the-art performance in terms of detection accuracy, recall, and F1 score. Furthermore, we show that ExHalder can generate high-quality natural language explanations to justify its prediction results.

Contributions. To summarize, our major contributions include: (1) a novel framework that automatically detects news headline hallucinations with limited manually labeled data; (2) an effective method for integrating natural language explanations into the detection pipeline and enabling the model to generate human-readable explanation; (3) a real-world headline hallucination detection datasets curated by news-domain experts; and (4) extensive experiments on seven real-world datasets that verify both the hallucination detection accuracy and the generated explanation quality.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 formalizes our problem. Then, we present our ExHalder framework in Section 4 and conduct experiments in Section 5. Finally, we conclude this paper in Section 6.

2 RELATED WORK

News Headline Generation. Automated news headline generation, widely considered as a special form of document summarization task, aims to generate a headline-style summary from either a single news article [2, 40] or a set of news articles related to the same event [23]. Early studies address this task by adopting an extractive approach [31, 39] that first selects words from the article and then organizes them into the output headline via statistical models [3, 18, 50]. This approach achieves limited success as some extracted words are incoherent [1] and the traditional statistical models lack expressive powers to generate vivid text. Recently, the advances of natural language generation research [47, 54, 64] lead to more abstractive headline generation methods [2, 6, 22, 23, 33]. They adopt the encoder-decoder architecture [12, 52] where the encoder synthesizes the knowledge in the news article(s) using vector representations and the decoder outputs the generated headline in a word-by-word fashion with potential constraints (e.g., length control [30], keyword preservation [37], or style preference [61]). Although the overall quality improvements have been made, people observe that these generation models often will output hallucinated headlines that are not supported by the underlying news articles [29, 60]. This hallucination issue becomes a key blocker for deploying web-scale automated headline generation models in industry, which motivates us to study the news headline hallucination detection problem in this work.

Hallucination Detection. Recent years have witnessed the great improvements of many natural language generation (NLG) models. One remaining challenge for deploying these NLG models in real-world systems is the hallucination issue that refers to the scenario where the generated content being nonsensical or unfaithful to the provided source content [20, 29, 36]. Many studies propose to mitigate the hallucination issue by either cleaning the model training data [21, 46] or learning a classifier to postprocess/filter

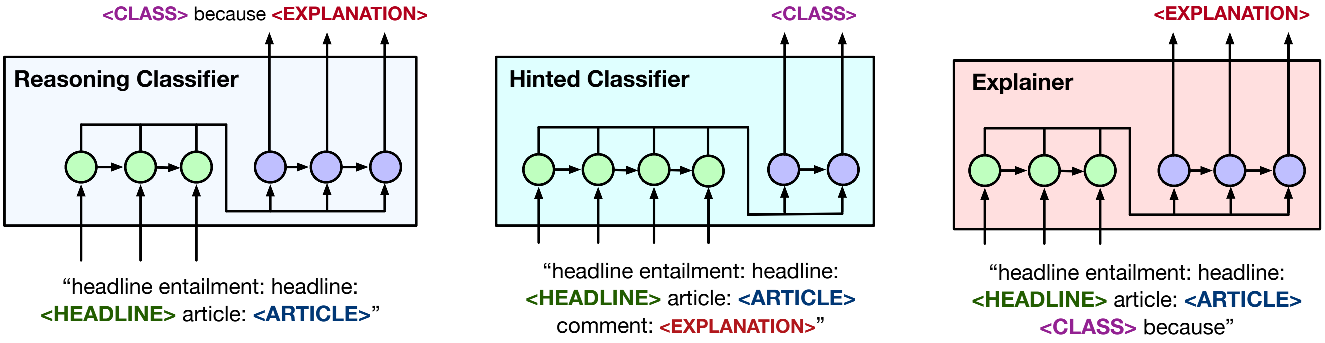

<CLASS> because <EXPLANATION>
Reasoning Classifier
"headline entailment: headline: <HEADLINE> article: <ARTICLE>"
<CLASS>
Hinted Classifier
"headline entailment: headline: <HEADLINE> article: <ARTICLE> comment: <EXPLANATION>"
<EXPLANATION>
Explainer
"headline entailment: headline: <HEADLINE> article: <ARTICLE> <CLASS> because"

Figure 2: Key Components of ExHalder framework.

generated contents [8, 10, 45]. In a boarder sense, our study falls into the second category and further enhances the classifier with a natural language explanation component.

Natural Language Inference. The task of natural language inference [5] (also called textual entailment [4, 16]) aims to predict if a given "premise" text entails, contradicts, or is neutral with regard to another "hypothesis" text. As this task can measure the model's language reasoning capability and has multiple large datasets [5, 58, 63], there have been studies on how to adapt it for other language tasks such as weakly-supervised classification [51], sentence embedding learning [14], and fact checking [48]. Among these studies, the most relevant are those utilizing trained NLI models for measuring the faithfulness of summarization methods [38, 43, 55]. However, different from this work, they do not leverage the explanation information. In contrast, our experiments show that these explanations can help better transfer the knowledge from the NLI task to the headline hallucination detection task.

Natural Language Explanation. Leveraging natural language explanations to improve machine learning models has long been studied in the literature. Typical usages include feeding the human-written explanations as additional input signals [25, 34] or treating them as model outputs and training the model to reproduce them [41]. Although how models benefit from these explanations still remains an active research problem [24], the general finding is that these natural language explanations could be particularly useful when only limited amount of labeled data are provided [32, 57, 62]. In this work, we study how to effectively leverage these explanations to enhance the hallucination detection accuracy and explore the possibility of generating free-text explanations to justify model's reasoning rationale.

3 PROBLEM FORMULATION

In this section, we first introduce the notations used later in the paper and then present our problem formulation.

Notations. A news *article* $\mathbf{d} \in \mathcal{D}$ is a document composed of a token sequence $[d_1, d_2, \dots]$. A news *headline* $\mathbf{h} \in \mathcal{H}$ is a succinct summary of the news article, represented by another token sequence $[h_1, h_2, \dots]$. Although the news headline is typically generated based on the news article, those generation methods sometimes encounter the *hallucination* issue, namely the generated headline is not entailed by its corresponding news article. Given a pair of article and headline $\langle \mathbf{d}_i, \mathbf{h}_i \rangle$, we use $s_i \in \mathcal{S} = \{0, 1\}$ to indicate if the headline is supported by the article or not. Optionally, we may have a natural language explanation to elucidate why the article supports or contradicts the headline. We use another token sequence $\mathbf{e} = [e_1, e_2, \dots]$ to denote this free-text explanation.

Problem Definition. The task of *news headline hallucination detection* is to learn a predictor $\mathbf{f}(\cdot) : \mathcal{D} \times \mathcal{H} \rightarrow \mathcal{Y}$ that takes a pair of $\langle$news article, news headline$\rangle$ as the input and predicts if the news headline is supported by the news article. Based on the available resources for learning the predictor $\mathbf{f}(\cdot)$, we further consider two settings: (1) **supervised setting** where we have a small set of N labeled examples $\{\mathbf{d}_i, \mathbf{h}_i, s_i\}|_{i=1}^{N}$ to train the predictor, and (2) **zero-shot setting** where we do not have any labeled example and have to exploit knowledge from other related tasks.

4 EXHALDER: EXPLANATION-ENHANCED HEADLINE HALLUCINATION DETECTOR

In this section, we first introduce three key components of our ExHalder framework. Then, we elaborate on how ExHalder utilizes these components for news headline hallucination detection and how to train the ExHalder framework. Finally, we discuss the inference procedure of ExHalder framework.

4.1 Key Components of ExHalder Framework

In this work, we adopt the widely used encoder-decoder architecture [12, 52] due to its strong representation power and wide applicability for both classification and generation tasks. The encoder first compresses the information of an input sequence $\mathbf{x} = [x_1, x_2, \dots]$ into its vector representation and then the decoder generates tokens in the output sequence $\mathbf{y} = [y_1, y_2, \dots]$ one at a time. Specifically, given the full input sequence $\mathbf{x}$ and the output sequence prefix $\mathbf{y}_{1:i-1} = [y_1, y_2, \dots, y_{i-1}]$, we produce the token y_i as follow:

$$\mathbf{P}(y_i|\mathbf{x}, \mathbf{y}_{1:i-1}) = \frac{\exp(\mathbf{v}_i \cdot \mathbf{E}(y_i))}{\sum_{y' \in \mathcal{V}} \exp(\mathbf{v}_i) \cdot \mathbf{E}(y'))}, \quad (1)$$

where $\mathbf{v}_i$ is the decoder output hidden vector corresponding to token y_i, $\mathbf{E}(y)$ is the embedding of a token y, and $\mathcal{V}$ denotes the entire vocabulary.

Although initially proposed for generation tasks, the encoder-decoder models can also be applied to classification problems by (1)

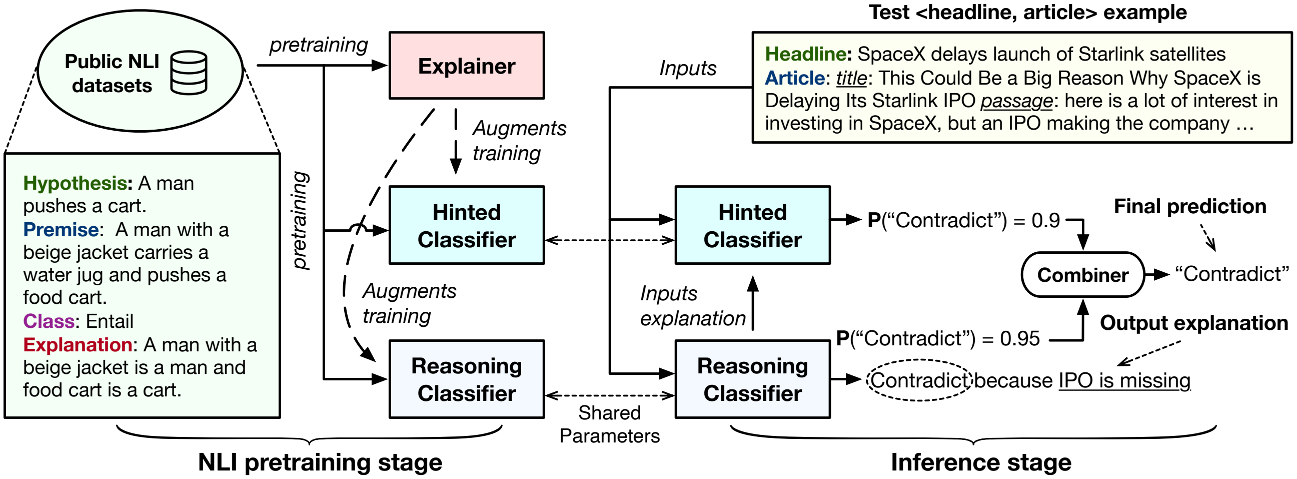

Public NLI datasets
pretraining
Explainer
Augments training
Hinted Classifier
pretraining
Augments training
Reasoning Classifier
Hypothesis: A man pushes a cart.
Premise: A man with a beige jacket carries a water jug and pushes a food cart.
Class: Entail
Explanation: A man with a beige jacket is a man and food cart is a cart.
Shared Parameters
Test <headline, article> example
Headline: SpaceX delays launch of Starlink satellites
Article: title: This Could Be a Big Reason Why SpaceX is Delaying Its Starlink IPO passage: here is a lot of interest in investing in SpaceX, but an IPO making the company …
Inputs
Hinted Classifier
P("Contradict") = 0.9
Final prediction
Combiner
"Contradict"
Inputs explanation
Output explanation
P("Contradict") = 0.95
Reasoning Classifier
Contradict because IPO is missing
NLI pretraining stage
Inference stage

Figure 3: The ExHalder framework overview.

choosing one special token for each possible class; (2) forcing the model to do one-step decoding; and (3) mapping the output token y_1 to its corresponding class as the final prediction [44]. Take our hallucination detection task as an example. We can use the special token 'c' in vocabulary $\mathcal{V}$ to represent the "contradictary" class and compute the hallucination probability as $\mathbf{P}(y_1 = `c' | \mathbf{x})$.

In the ExHalder framework, given a pair of ⟨article, headline⟩ ($\langle \mathbf{d}_i, \mathbf{h}_i \rangle$) along with its class label s_i and label explanation $\mathbf{e}_i$, we define the following three components based on how we construct their input sequences $\mathbf{x}$ and expect the output sequences $\mathbf{y}$ will be. Figure 2 shows an architecture overview of these three components.

Reasoning Classifier. The input sequence $\mathbf{x}$ is of format "headline entailment: headline: <HEADLINE> article: <ARTICLE>" where the "<HEADLINE>" and "<ARTICLE>" are two placeholders and will later be replaced with the contents in the news headline $\mathbf{h}$ and the news article $\mathbf{d}$. The output sequence $\mathbf{y}$ is of format "<CLASS> because <EXPLANATION>" where the placeholder token "<CLASS>" (one of {"Entail, "Contradict"}) indicates if the news article entails or contradicts the news headline, and the placeholder token "<EXPLANATION>" corresponds to the natural language explanation $\mathbf{e}$. When the rater does not provide any explanation for the labeled example during the curation process, this "<EXPLANATION>" token could simply be an empty string. Note here if we throw away the "because <EXPLANATION>" part in the output sequence $\mathbf{y}$, the reasoning classifier will degenerate into a standard classifier with the encoder-decoder architecture.

Hinted Classifier. As its name suggests, the input of hinted classifier goes beyond the one used for the reasoning classifier and includes the natural language explanation as the "hint". Specifically, we append a string "comment: <EXPLANATION>" after the reasoning classifier's input and teach the model to output a single token "<CLASS>" to indicate the final predicted class. The hinted classifier is expected to achieve better classification performance than the reasoning classifier because (1) its input contains more signals from the additional "comment: <EXPLANATION>" part, and (2) it does not waste representative power for the explanation generation.

Explainer. Different from the previous two "classifiers", the explainer inputs a sequence that already contains the class information and aims to output a natural language sentence to explain this class. Specifically, the input sequence of the explainer is of format "headline entailment: headline: <HEADLINE> article: <ARTICLE> <CLASS> because" and the output sequence will be just the natural language explanation itself.

4.2 The ExHalder Framework

Our ExHalder framework is built upon the above three key components for news headline hallucination detection. As both the reasoning classifier and the hinted classifier contain the prediction result "<CLASS>" in their outputs, one may argue that we can directly adopt supervised learning techniques to train these two classifiers for hallucination detection. This approach, however, requires massive labeled data which are often inaccessible for real-world applications. Therefore, in this work, we propose two novel techniques to address such a label data scarcity issue: (1) pretraining with large-scale natural language inference (NLI) datasets, and (2) augmented training with human-written explanations. Figure 3 shows an overview of our ExHalder framework.

4.2.1 NLI-based Pretraining. The natural language inference (NLI) task aims to predict if a given "hypothesis" is supported/entailed by another input "premise" text. Take the case in Figure 3 as an example, the hypothesis "*A man pushes a cart*" is supported by the premise "*A man with a beige jacket carries a water jug and pushes a food cart.*" and thus the target class is "Entail". We observe that this NLI task shares many similarities with our news headline hallucination detection (HHD) task. Both of them aim to detect if one piece of text ("headline" in the HHD task and "hypothesis" in the NLI task) is supported/entailed/grounded by another piece of text ("article" in the HHD task and "premise" in the NLI task). Such a connection enables us to transfer knowledge from the NLI task to our news domain HHD task. Furthermore, different from the case in the news domain HHD with limited labeled data, there are many large-scale publicly available NLI datasets [5, 7, 42, 58, 63].

Based on the above observation, in this work, we propose to pretrain all the components in ExHalder using the NLI datasets. Specifically, we use the eSNLI [7] and ANLI [42] datasets for pretraining as they both contain human written natural language explanations. Given a NLI example ⟨hypothesis, premise, label⟩, we first construct one training example by replacing the "<HEADLINE>" and the "<ARTICLE>" placeholder tokens with the "hypothesis" text and the "premise" text, respectively. Then, we train our reasoning classifier, hinted classifier, and explainer models using the standard teacher-forcing technique [59].

4.2.2 Explainer-augmented Training. Due to language variability, people have different ways to express the same underlying rationale. However, in the existing NLI datasets, due to constrained manual curation resources, each example has only a very limited amount of human-written explanation(s) (e.g., 1 for the eSNLI dataset and 1-3 for the ANLI dataset). To obtain more explanations and use them to train the hinted classifier and the reasoning classifier, we propose to augment the existing NLI datasets with a learned explainer. Specifically, after the initial pretraining stage, we use the learned explainer to generate K additional explanations for each NLI example. Then, we merge these augmented examples with the examples in the original NLI dataset and continue to train the hinted classifier and reasoning classifier with this augmented dataset. More training details are discussed in the experiment section.

4.2.3 Optional Domain Fine-tuning. For both the NLI-based pretraining step and the explainer-augmented training step, we only use the general domain datasets. When additional news domain-specific datasets are available, we can follow the same procedure above and further fine-tune the components in our ExHalder framework. In this work, we collect a new headline hallucination dataset and perform this domain fine-tuning step in one of our experiment settings (c.f. Section 5.1).

4.3 ExHalder Inference

At the inference stage, we are given a test ⟨article, headline⟩ pair and apply the learned hinted classifier and reasoning classifier to make a prediction. Specifically, we first feed the test example into the reasoning classifier and parse its output sequence into the predicted class and the explanation sentence. Then, we concatenate this generated explanation with the original headline and article and treat it as the input sequence of the hinted classifier. We use the hinted classifier to obtain another class prediction. Finally, we use a combiner to aggregate the predictions from the reasoning classifier and the hinted classifier. Here, without requiring more labeled examples, we adopt a simple averaging strategy for the combiner. Namely, we average the probability scores from the reasoning classifier and the hinted classifier and return this averaged score as the final prediction probability[1].

5 EXPERIMENTS

In this section, we study the performance of ExHalder on two settings: (1) *supervised setting* where we have a small set of labeled ⟨article, headline⟩ pairs for model learning, and (2) *zero-shot setting* where no labeled data is provided.

5.1 News Headline Hallucination Detection with Supervision

5.1.1 Dataset. To the best of our knowledge, there is no publicly available news headline hallucination detection dataset. Therefore, in this paper, we collect a new dataset that contains 6270 human curated examples: 5190 examples for training, 349 examples for validation, and 731 examples for testing. Each example includes a triplet of ⟨news article, news headline, hallucination label⟩ where the headline is generated from NHNet [23] and the label is obtained from multiple human experts according to a common guideline. Specifically, we ask three full-time journalism degree holders in the news domain to rate each example and determine the final hallucination label through majority voting. Among these examples, 1934 of them are labeled as "hallucinated" and the remaining 4336 examples are labeled as "entailed". Furthermore, there are 2074 examples with additional rater-written comments (besides binary hallucination labels) and we treat them as user-provided explanations. The dataset is publicly available at: `https://bit.ly/exhalder-dataset`.

5.1.2 Compared Methods. We compare the following methods for the headline hallucination detection task:

- SVM [15]: We manually extract a set of features based on the textual string of the news headline and the news article (e.g., their corresponding sequence lengths, the number of overlapping words, some word-level editing distances like Jaro-Winkler distance [13], *etc.*), and train a standard SVM model with the RBF kernel for predictions.
- XGBoost [11]: Similar to the above SVM method, we feed those handcrafted features to the standard XGBoost classification model for detecting the hallucinations.
- BERT$_{base}$ [17]: We concatenate the headline and the article text (with a [SEP] separator) and feed it into the pretrained BERT base model for prediction.
- T5$_{xxl}$ [44]: Similar to BERT, we input the concatenated headline and article to the encoder module of T5 and use its decoder to output one single token indicating the final predicted class.
- T5$_{xxl}$ + Exp: We incorporate the natural language explanation information into the T5$_{xxl}$ model by requiring its decoder to output the class token followed by the explanation. This is similar to the reasoning classifier architecture in our ExHalder framework.
- ExHalder-NoPT: Our ExHalder framework without the NLI-based pretraining step. Namely, we just train the explainer, the reasoning classifier, and the hinted classifier on our news-domain hallucination detection training set.
- ExHalder-NoEX: Our ExHalder framework with the NLI-based pretraining step but without leveraging any explanation information. Namely, we force the reasoning classifier to just output one token indicating the hallucination label and remove the hinted classifier as well as the explainer components.

[1] When more labeled examples are available, another combiner design is to train a small model to calibrate and aggregate the probability scores from both the reasoning classifier and the hinted classifier.

Methods	Accuracy	Precision	Recall	F1
SVM	57.31	28.65	20.50	23.90
XGBoost	60.19	42.39	60.67	49.91
BERT$_{base}$	73.46	71.43	31.38	43.60
T5$_{xxl}$	82.39	76.29	66.93	71.29
T5$_{xxl}$ + Exp	82.62	78.98	64.15	70.63
ExHalder-NoPT	82.08	75.96	66.11	70.69
ExHalder-NoEX	83.17	80.01*	64.71	71.54
ExHalder-NoHC	84.08*	82.06*	65.69	72.96*
ExHalder	**84.46***	**82.63***	**67.16***	**74.08***

Table 1: Quantitative results on the news headline hallucination detection dataset. The superscript * means the improvement is statistically significant compared to T5$_{xxl}$.

- ExHalder-NoHC: Our ExHalder framework without the hinted classifier module. Namely, we only train the explainer and the reasoning classifier during the pretraining stage and use the reasoning classifier alone for prediction at the inference stage.
- ExHalder: The full version of our proposed framework.

We implement SVM using scikit-learn, XGBoost using its official codebase[2], and BERT$_{base}$ method using the Tensorflow Model Garden[3]. For T5$_{xxl}$, T5$_{xxl}$+Exp, and ExHalder along with its variants, we develop them based on the T5X library[4] and use the T5-11B model in the following experiments. More implementation details and hyper-parameter settings are discussed in Appendix A.

5.1.3 Experiment Settings. As we formulate the headline hallucination detection as a classification problem, we adopt the standard classification evaluation metrics: Accuracy, Precision, Recall, and F1 score. Among these metric, we emphasize that the recall value indicates the percentage of hallucinated headlines captured by the hallucination detector. Better recall means less misleading headlines will be surfaced to users and thus leads to more positive user experiences. For each tested method, we run it for five times and report the averaged results. Finally, for performance comparisons, we we conduct statistical significance test using the two-tailed paired t-test with 95% confidence level.

5.1.4 Experiment Results. Below we first present the main experiment results and compare ExHalder with the baseline methods. Then, we conduct ablation analysis to study how the key components of ExHalder impact the framework overall performance. Finally, we present a few case studies to demonstrate the potential impacts of ExHalder in real-world scenarios.

1. Overall Detection Performance. Table 1 presents the results of all compared methods. First, we can see that the results of those traditional methods with manual feature engineering (i.e., SVM, XGBoost) are unsatisfactory. This shows that headline hallucination detection is a challenging task and requires models to capture the subtle semantic differences between the article and the headline.

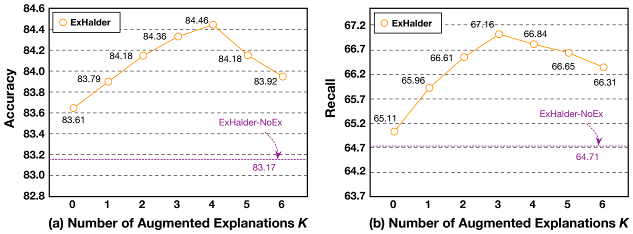


Figure 4: Parameter sensitivity analysis on the news hallucination detection dataset. We vary the number of explanations generated by the explainer component and compute the accuracy and recall of ExHalder.

Second, we compare ExHalder with ExHalder-NoPT and see that the NLI-based pretraining indeed helps us to better identify the hallucinated headlines by warming up the model with entailment task semantics. Third, by comparing ExHalder with ExHalder-NoEX, we observe further performance improvements and this demonstrates that injecting the explanation information into the model training process is useful. Finally, we can see our proposed ExHalder has the overall best performance across all the metrics and defeats the second-best method by a large margin.

2. Ablation Analysis of Model Components. ExHalder contains three key components: a reasoning classifier, a hinted classifier, and an explainer. The above ExHalder-NoEX demonstrates the importance of the reasoning classifier component and the explainer component. Here, we study how the hinted classifier components affect the performance of ExHalder. As shown in Table 1, we can see that removing the hinted classifier leads to low prediction accuracy and significantly hurts the hallucination detection recall.

3. Explainer Augmentation Analysis. We continue to evaluate the explainer component by directly varying its parameter K, namely the number of its generated explanations used for augmenting reasoning and hinted classifier training. As shown in Figure 4, the model performance first increases as K increases until it reaches about 3 to 4 and then starts decreasing. Notably, the performance dropping rates vary across different evaluation metrics. The model accuracy drops faster compared to its recall. This is probably because the quality of generated explanations will decrease if we force the explainer to generate lots of explanations. Finally, we can see that for a wide range of K, the performance of ExHalder is better than ExHalder-NoEX, which further demonstrates the usefulness of free-text explanations.

4. Case Studies. Table 2 shows some ExHalder output examples. More case studies are presented in Appendix C. First, we observe that ExHalder can generate high-quality human-readable explanations to justify its prediction. In the first example, the model output explanation "conflicting dates - 2021 vs 2019." captures the key difference between the headline and the article, and closely resembles the human written explanation "the date in the headline different from the one appearing in the article".

Second, we can see that ExHalder is able to help us identify potential labeling errors. Take the second example as one case, the rater mistakenly labels it as an "Entail" case but in fact it should be misleading because the headline suggests the Starlink satellites launch is delayed but the article is about the delay of Starlink IPO.

[2] `https://xgboost.readthedocs.io/en/stable/index.html`.

[3] `https://github.com/tensorflow/models/tree/master/official/nlp`.

[4] `https://github.com/google-research/t5x`.

Headline: WWE SmackDown results - 2/9/19
Article: *title*: WWE Friday Night SmackDown Results (3/26/21) *passage*: WWE Friday Night SmackDown Results March 26, 2021 Report by Lovell Porter for Wrestlezone.com You can also participate via social media by using the #WZChat hashtag to voice your thoughts on tonight's show. We want you to share our exclusive coverage page ...
Human rated class: "Contradict"
Human provided explanation: "the date in the headline different from the one appearing in the article."
ExHalder output: "Contradict because conflicting dates - 2021 vs 2019."

Headline: SpaceX delays launch of Starlink satellites
Article: *title*: This Could Be a Big Reason Why SpaceX Is Delaying Its Starlink IPO *passage*: here is a lot of interest in investing in SpaceX, but an IPO making the company available to everyday investors is likely still a long way off. CEO Elon Musk has said SpaceX internet service provider subsidiary Starlink is much closer to ...
Human rated class: "Entail"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder output: "Contradict because IPO is missing in the headline which makes it misleading."

Headline: Lockdown at 16 hotspots in Thane till March 31
Article: *title*: Maharashtra: Thane Municipal Corporation does U-turn on lockdown in 16 hotspots *passage*: HANE: The Thane Municipal Corporation (TMC) in Maharashtra, on Tuesday evening, did a complete U-turn on its Monday announcement of lockdown in 16 hotspots in the district. The TMC issued fresh guidelines tapering restrictions to the respective floors of a building or wings depending on the caseload, till March end, officials said....
Human rated class: "Entail"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder output explanation: "Contradict because till March 31 is unsupported."

Table 2: ExHalder output case studies. We use two colors to highlight the key information in headlines and articles.

Headline: Kyle takes his own life in Hollyoaks
Article: *title*: Does Kyle die in Hollyoaks? *passage*: Hollyoaks will tackle the subject of male suicide this week though a hard-hitting storyline featuring Kyle Kelly (Adam Rickitt). ... However, last week, conflict ensued between the friends after Kyle caused a devastating car crash while on drugs ...
Human rated class: "Contradict"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder explainer output 1: "Kyle is in a car crash so he doesn't take his own life."
ExHalder explainer output 2: "Kyle is in a car crash, not taking his own life."
ExHalder explainer output 3: "Kyle does not die in Hollyoaks. The show is about male suicide."

Table 3: Case study of the explainer module in our ExHalder framework.

We can capture this error based on the model output explanation "IPO is missing in the headline which makes it misleading".

Moreover, we can see that the generated explanation enables us to understand why the model makes a certain mistake. As shown by the last example in Table 2, the headline is indeed supported by the news article but our ExHalder predicts it to be a contradiction case because the "till March 31" is not supported by the article. Diving into the news article, we can see the "till March 31" information is referred to as "till March end" in the middle part of the article's main passage. The model fails to recognize that the "March end" is synonymous with "March 31" and thus makes the wrong prediction. This observation can motivate researchers to later study how to further increase the hallucination accuracy by improving the model's temporal reasoning ability.

Finally, Table 3 shows one example where the original curator does not provide any rating explanations other than the binary class label while our ExHalder explainer component can successfully generate valid explanations for the rated class.

5.2 Zero-shot Hallucination Detection

5.2.1 Datasets. We further evaluate the zero-shot performance of ExHalder when no in-domain training data is provided. Specifically, we adopt the four summarization hallucination detection datasets: MNBM [38], FRANK [43], QAGS [56], SummEval [19] and two fact verification datasets: FEVER [55], Vitamin-C [49] in the TRUE benchmark [27]. Each dataset contains a set of ⟨target text, grounding text, hallucination label⟩ triplets where the binary label indicates if the target text is hallucinated based on the grounding text. In the following experiments, we treat the target text as the "headline" and the grounding text as the "article". More dataset details are available in Appendix B.

5.2.2 Compared Methods. We compare our ExHalder framework and its variant ExHalder-NoEX with two best-performing methods

Dataset: QAGS
Headline: The number of illegal immigrants detained or arrested at the uk's busiest port has more than doubled in a year, figures show.
Article: There has been a large increase in the number of illegal immigrants detained or arrested by police at dover. Some 563 were held or arrested by port of dover police officers last year - up from 148 in 2013, a bbc radio kent freedom of information request shows. ...
Human rated class: "Contradict"
ExHalder output: "Contradict because The number of illegal immigrants detained or arrested at dover - one of the uk's busiest ports - has more than doubled, from 148 to 563."

Dataset: Vitamin-C
Headline: Averaged over more than 20 critics , On the Basis of Sex scored less than 62 % .
Article: On Metacritic , the film has a weighted average score of 61 out of 100 , based on 21 critics , indicating "generally favorable reviews" .
Human rated class: "Entail"
ExHalder output: "Entail because The film scored 61 out of 100 , which is less than 62 % ."

Table 4: ExHalder output case studies on TRUE benchmark datasets. If the article and the headline is contradictory, we use two different colors to highlight the key differences. Otherwise, we use one single color to underscore the shared key information.

Datasets	Q2	ANLI	ExHalder-NoEX	ExHalder
MNBM [38]	66.5	66.7	73.8	**75.4**
FRANK [43]	82.9	**83.5**	81.3	83.3
QAGS [56]	78.3	75.3	76.6	**78.4**
SummEval [19]	77.3	72.9	85.6	**87.0**
FEVER [55]	82.7	**90.2**	87.4	88.3
Vitamin-C [49]	75.7	74.7	84.8	**85.1**
Average	77.23	77.21	81.58	**82.91**

Table 5: Accuracy results on the TRUE datasets [27].

in the original TRUE paper: (1) **ANLI** which, similar to our approach, first trains a T5-11B model using the ANLI dataset [42] and then directly applies the learned model to detect the hallucinations, and (2) **Q2** [28] which first uses a question generation module to generate questions with answer spans from the target text and then applies a question answering (QA) model on the grounding text to answer the above-generated questions. Finally, it computes the overlap between each true answer span and its corresponding QA model output answer span and outputs the final hallucination score.

5.2.3 Experiment Settings. As no training example is provided in the TRUE benchmark, we reuse the ExHalder checkpoint after the NLI-based pretraining step and directly conduct the inference step of ExHalder on all tested datasets. For fair comparisons, we follow the previous practices [27] to directly tune the binary cutoff threshold on the development set and report the best performance (in terms of accuracy) of all baseline methods in the original paper.

5.2.4 Experiment Results. Table 5 shows the overall results on all six evaluated datasets. We can see that both ExHalder and ExHalder-NoEX can outperform the previous best methods and our ExHalder framework achieves new state-of-the-art results. Moreover, by comparing ExHalder with ExHalder-NoEX, we observe that adding explanation information is particularly useful in the zero-shot transfer learning setting. We also demonstrate that ExHalder can augment the TRUE benchmark by providing interesting and insightful free-text explanations for the existing labels. As shown in Table 4, ExHalder generates high quality human-readable explanations to explain its prediction results. Take the case from the QAGS dataset as an example, ExHalder's output explanation captures the subtle semantic difference between "*uk's busiest port*" and "*dover, one of the uk's busiest ports*" and justifies why it makes the "Contradict" prediction. Similarly, in the example from Vitamin-C dataset, ExHalder reiterates the fact that "*scored 61 out of 100*" (from the news article) implies "*less than 62%*" (mentioned in the headline) and thus the headline is supported by the article. More case studies are presented in Appendix D.

6 CONCLUSIONS AND FUTURE WORK

This paper studies how to automatically detect news headline hallucinations with a limited amount of labeled data. We propose a novel ExHalder framework which adapts knowledge from public NLI datasets into the news domain and generates natural language explanations to justify its prediction results. Extensive experiments on one newly collected dataset and six public datasets demonstrate that ExHalder can accurately identify hallucinated news headlines along with high-quality human-readable explanations.

As a first-punch solution for detecting news headline hallucinations, we believe ExHalder can be improved in many ways. Interesting future directions include: (1) utilizing the validation set to learn a better combiner that better aggregates the predictions results from the reasoning classifier and the hinted classifier, (2) incorporating large language models (e.g., GPT-3, PaLM, ChatGPT) into ExHalder for better zero- and few-shot performance, (3) expanding the scope of ExHalder to the multilingual setting for detecting international news headline hallucinations, (4) formatting the ExHalder output explanations to increase their readability, (5) enforcing the ExHalder output explanation itself to be entailed by the original news article and headline, and (6) extending ExHalder to resolve multi-document headline hallucination problems where the headline is generated from multiple documents and we need to predict if it is hallucinated based on a whole set of documents.

REFERENCES

[1] Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. In *ACL*.

[2] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A Dataset and Generic Framework for Personalized News Headline Generation. In *ACL*.

[3] Michele Banko, Vibhu Mittal, and M. Witbrock. 2000. Headline Generation Based on Statistical Translation. In *ACL*.

[4] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

[6] Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of Transformer-Based Models for News Headline Generation. *ArXiv* abs/2007.05044 (2020).

[7] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *NeurIPS*.

[8] Mengyao Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual Error Correction for Abstractive Summarization Models. In *EMNLP*.

[9] Pew Research Center. 2021. *Newspapers Fact Sheet - Pew Research Center*. https://www.pewresearch.org/journalism/fact-sheet/newspapers/

[10] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection. In *NAACL*.

[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[12] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *EMNLP*.

[13] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *IIWeb*.

[14] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP*.

[15] Corinna Cortes and Vladimir Naumovich Vapnik. 2004. Support-Vector Networks. *Machine Learning* 20 (2004), 273–297.

[16] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *MLCW*.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[18] B. Dorr, David M. Zajic, and Richard M. Schwartz. 2003. Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In *HLT-NAACL 2003*.

[19] A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.

[20] Katja Filippova. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In *EMNLP Findings*.

[21] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating Training Corpora for NLG Micro-Planners. In *ACL*.

[22] Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-Attentive Model for Headline Generation. In *ECIR*.

[23] Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. *Proceedings of The Web Conference 2020* (2020).

[24] Peter Hase and Mohit Bansal. 2022. When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data. In *LNLS*.

[25] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?. In *EMNLP Findings*.

[26] Tatsuru Higurashi, Hayato Kobayashi, Takeshi Masuyama, and Kazuma Murao. 2018. Extractive Headline Generation Based on Learning to Rank for Community Question Answering. In *COLING*.

[27] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Y. Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *NAACL*.

[28] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. In *EMNLP*.

[29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *ArXiv* abs/2202.03629 (2022).

[30] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling Output Length in Neural Encoder-Decoders. In *EMNLP*.

[31] Nataliya Kochetkova, Ekaterina V. Pronoza, and Elena Yagunova. 2018. News Headline as a Form of News Text Compression. In *SocInfo*.

[32] Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Jamin Chen, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren. 2020. LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation. In *ACL*.

[33] Zhengpeng Li, Jian Wu, Jiawei Miao, and Xinmiao Yu. 2022. News headline generation based on improved decoder from transformer. *Scientific Reports* 12 (2022).

[34] Weixin Liang, James Y. Zou, and Zhou Yu. 2020. ALICE: Active Learning with Contrastive Natural Language Explanations. In *EMNLP*.

[35] Bill MacCartney. 2009. *Natural language inference*. Stanford University.

[36] Amr Magdy and Nayer M. Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*.

[37] Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained Abstractive Summarization: Preserving Factual Consistency with Constrained Generation. *ArXiv* abs/2010.12723 (2020).

[38] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*.

[39] Hadi Mohammadzadeh, Thomas Gottron, Franz Schweiggert, and Gerhard Heyer. 2012. TitleFinder: extracting the headline of news web pages based on cosine similarity and overlap scoring similarity. In *WIDM ’12*.

[40] Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A Case Study on Neural Headline Generation for Editing Support. In *NAACL*.

[41] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *ArXiv* abs/2004.14546 (2020).

[42] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL*.

[43] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *NAACL*.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[45] Hannah Rashkin, D. Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In *ACL*.

[46] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2021. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery* 36 (2021), 318–354.

[47] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*.

[48] Aalok Sathe and Joonsuk Park. 2021. Automatic Fact-Checking with Document-level Annotations using BERT and Multiple Instance Learning. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)* (2021).

[49] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *NAACL*.

[50] D. Zajic R. Schwartz, Blanche E. Door, and Richard M. Schwartz. 2002. Automatic Headline Generation for Newspaper Stories.

[51] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *NAACL*.

[52] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*.

[53] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach. In *IJCAI*.

[54] Cheng Tang, Frank Guerin, Yucheng Li, and Chenghua Lin. 2022. Recent Advances in Neural Text Generation: A Task-Agnostic Survey. *ArXiv* abs/2203.03047 (2022).

[55] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. *ArXiv* abs/1811.10971 (2018).

[56] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *ACL*.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv* abs/2201.11903 (2022).

[58] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*.

[59] Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation* 1 (1989), 270–280.

[60] Yijun Xiao and William Yang Wang. 2021. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *EACL*.

[61] Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. In *EMNLP*.

[62] Qinyuan Ye, Xiao Huang, and Xiang Ren. 2020. Teaching Machine Comprehension with Compositional Explanations. In *EMNLP Findings*.

[63] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. In *NAACL*.

[64] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*.

A EXPERIMENT DETAILS ON NEWS HALLUCINATION DETECTION DATASET

For all compared methods, we tune their hyper-parameters using the validation set, select the best ones, and report the corresponding results on the test set. Specifically, we have: for SVM[5], we use the RBF kernel with C=0.1 and degree=4; for XGBoost[6], we select gamma=1.0, max_depth=3, min_child_weight=1, subsample=1.0, colsample_bytree=0.5, and n_estimators=30; for BERT$_{base}$[7] and T5$_{xxl}$[8] methods, we select batch_size=64 and learning_rate=1e-3. Both methods use a constant learning rate scheduler and are trained for 10k steps with 1k warmup steps. For our ExHalder framework and its variants, during the NLI-based pretraining stage, we choose batch_size=128, constant learning_rate=1e-3, and the number of explainer generated explanations $K = 1$. During the domain fine-tuning stage, we select batch_size=64, constant learning_rate=1e-3, and the number of explainer generated explanations $K = 3$. Both stages are trained for 10k steps with 1k warmup steps. Finally, we train BERT$_{base}$, T5$_{xxl}$, and our models on TPU v3.

B TRUE BENCHMARK DATASETS STATISTICS

Table 6 lists the statistics of TRUE benchmark datasets.

Datasets	# of Examples
MNBM [38]	2,500
FRANK [43]	671
QAGS [56]	239
SummEval [19]	1,600
FEVER [55]	18,209
Vitamin-C [49]	63,054

Table 6: Statistics of TRUE Benchmark Datasets.

C CASE STUDIES ON NEWS HALLUCINATION DETECTION DATASET

Table 7 lists case studies on our news hallucination detection dataset.

D CASE STUDIES ON TRUE BENCHMARK DATASETS

Table 8 lists case studies on TRUE benchmark datasets.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[6] https://xgboost.readthedocs.io/en/stable/index.html

[7] https://github.com/tensorflow/models/tree/master/official/nlp

[8] https://github.com/google-research/t5x

Headline: WWE SmackDown results - 2/9/19
Article: *title*: WWE Friday Night SmackDown Results (3/26/21) *passage*: WWE Friday Night SmackDown Results March 26, 2021 Report by Lovell Porter for Wrestlezone.com You can also participate via social media by using the #WZChat hashtag to voice your thoughts on tonight's show. We want you to share our exclusive coverage page ...
Human rated class: "Contradict"
Human provided explanation: "the date in the headline different from the one appearing in the article."
ExHalder output: "Contradict because conflicting dates - 2021 vs 2019."

Headline: SpaceX delays launch of Starlink satellites
Article: *title*: This Could Be a Big Reason Why SpaceX Is Delaying Its Starlink IPO *passage*: here is a lot of interest in investing in SpaceX, but an IPO making the company available to everyday investors is likely still a long way off. CEO Elon Musk has said SpaceX internet service provider subsidiary Starlink is much closer to ...
Human rated class: "Entail"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder output: "Contradict because IPO is missing in the headline which makes it misleading."

Headline: OnePlus 8T Cyberpunk 2077 Edition launched
Article: *title*: OnePlus 8T Cyberpunk 2077 Edition unveiled *passage*: OnePlus has unveiled a new special edition smartphone, the OnePlus 8T Cyberpunk 2077 Edition and the handset is now available to pre-order in China. The new OnePlus 8T Cyberpunk 2077 Edition ...
Human rated class: "Entail"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder output: "Contradict because launched VS unveiled."

Headline: Lockdown at 16 hotspots in Thane till March 31
Article: *title*: Maharashtra: Thane Municipal Corporation does U-turn on lockdown in 16 hotspots *passage*: HANE: The Thane Municipal Corporation (TMC) in Maharashtra, on Tuesday evening, did a complete U-turn on its Monday announcement of lockdown in 16 hotspots in the district. The TMC issued fresh guidelines tapering restrictions to the respective floors of a building or wings depending on the caseload, till March end, officials said....
Human rated class: "Entail"
Human provided explanation: "" (i.e., rater does not give any explanation)
ExHalder output explanation: "Contradict because till March 31 is unsupported."

Table 7: ExHalder output case studies. We use two colors to highlight the key information in headlines and articles.

Dataset: MNBM
Headline: singer rod temperton, who wrote the hit album michael jackson, has died at the age of 89.
Article: Temperton died in London last week at the age of 66 after "a brief aggressive battle with cancer", Jon Platt of Warner/Chappell music publishing said. Temperton's other hits included Off The Wall and Baby Be Mine for Jackson and Boogie Nights for his band Heatwave. Chic guitarist Nile Rodgers was among those paying tribute, tweeting: ...
Human rated class: "Contradict"
ExHalder output: "Contradict because he did die at the age of 66, not 89."

Dataset: FEVER
Headline: Heroes had a season with over 20 episodes.
Article: Heroes is an American psychological thriller science fiction television drama series created by Tim Kring that appeared on NBC for four seasons from September 25 , 2006 through February 8 , 2010 . The critically acclaimed first season had a run of 23 episodes and garnered an average of 14.3 million viewers in the United States , ...
Human rated class: "Entail"
ExHalder output: "Entail because The first season had a run of 23 episodes."

Table 8: ExHalder output case studies on TRUE benchmark datasets. If the article and the headline is contradictory, we use two different colors to highlight the key differences. Otherwise, we use one single color to underscore the shared key information.